KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Politechnika Wrocławska

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

ROZWÓJ POTENCJAŁU I OFERTY DYDAKTYCZNEJ POLITECHNIKI WROCŁAWSKIEJ

Wrocław University of Technology

Nanoengineering

Arkadiusz Wójs

COMPUTER MODELING OF SEMICONDUCTOR NANOSTRUCTURES

Wrocław 2011

Projekt współfinansowany ze środków Unii Europejskiej w ramach
Europejskiego Funduszu Społecznego

Wrocław University of Technology

Nanoengineering

Arkadiusz Wójs

COMPUTER MODELING OF SEMICONDUCTOR NANOSTRUCTURES

\

Wrocław 2011

Reviewer: Włodziemierz Salejda

ISBN 978-83-62098-50-7

Published by PRINTPAP Łódź, www.printpap.pl

Contents

I. Basic properties of nanostructures

Nanostructures are solid state (crystalline) objects of ultra-small sizes which fall intermediate between molecular and microscopic (micrometer) length scales. They are sufficiently small to exhibit quantization of dynamics of confined carriers (conduction electrons and/or valence holes). To allow control over the (small) number of confined carriers, nanostructures are usually made of semiconductors (→ semiconductor nanostructures). Typically small effective mass in the semiconductor medium (relative to vacuum) ensures persistence of quantum effects in nanostructures of typical length scales exceeding atoms or molecules by at least an order of magnitude (i.e., reaching tens of nanometers) – of course, at least at sufficiently low temperatures.

Nanostructures are most obviously differentiated with regard to the number of nanotextured spatial dimensions. In unconfined/untextured systems the electron motion remains continuous (i.e., not quantized) in all three spatial dimensions (3D). These are called *bulk* systems, to be contrasted with all kinds of nanostructures.

In the so-called *quasi-2D* nanostructures, quantum confinement occurs in a single spatial direction, leaving unconfined carrier motion over the remaining two-dimensional surface. Such structures involve surfaces, interfaces, and their combinations. For example, a *quantum well* is obtained between a pair of parallel interfaces, i.e., inside a thin layer of one semiconductor material sandwiched between the adjacent layers of another semiconductor, whose higher energy position of the conduction band (and/or lower energy position of the valence band) produce the potential energy barriers and so define a potential well inside. A natural, atomically 2D system is a single sheet of graphite (carbon) called graphene.

In the so-called *quasi-1D* nanostructures, confinement in two spatial dimensions leads to the quantized dynamics, and the continuous motion is only possible along a line. Among these systems one can list various forms of quantum wires, nanopillars, nanorods, and nanorings (closed loops). Atomic quasi-1D systems include carbon nanotubes (sufficiently narrow for strong quantization of the motion around the tube) and certain linear polymers.

Quasi-0D systems are often called *quantum dots* or *artificial atoms*, as they – similarly to their natural analogs – are nanoscale dot-like objects, containing a small and controllable number of confined electrons (and/or holes). Depending on the form or fabrication method, various quantum dots are also called nanocrystals, nanoparticles, nanoislands, or nanospheres.

1. Sizes, shapes, and materials

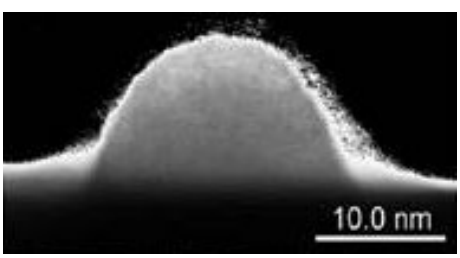
Depending on the fabrication process, quantum dots can have much greater variety of sizes and shapes than natural atoms or molecules. An obvious lower limit is the size of a single atom (consider the so-called 'covalent' radii of several atoms: H: 0.025 nm, C: 0.070 nm, Si: 0.110 nm, Ga: 0.130 nm, As: 0.115 nm), or the lattice constant of the involved semiconductor (Si-diamond: 0.543 nm, GaAs-zincblende: 0.565 nm, CdTe-zincblende: 0.648 nm). A dot

extending a mere 5-10 atoms will therefore measure about 3-6 nm in that given direction. The largest systems still called quantum dots measure tens of nm in some direction, corresponding to well over a hundred atoms.

In terms of shape, especially the larger quantum dots can have quite arbitrary geometry, defined by – to give but several examples – (i) remote electrodes, made to flexible shape by means of nanolithography and producing an electrostatic lateral confining potential for the carriers, initially trapped in two or less dimensions by some other method; (ii) physical shape of the nanoscale crystalline object either placed in vacuum/glass or inside another semiconductor characterized by an unfavorable band energy.

Of course, in smaller dots the actual shape of the electronic wave functions inside the dot does not accurately follow the shape of the physical nanostructure (or the confining potential), but instead it becomes much smoother, often nearly circular/spherical despite evidently broken rotational symmetry of the structure. Still, at least the relation between two or three spatial dimensions can be varied freely even in the smallest structures. For example, the so-called self-assembled dots can be grown in form of cut-out spheres or pyramids, with the aspect ratio (between the characteristic dimension in the plane of growth and normal to this plane) strongly dependent on the chosen materials (e.g., the relative lattice mismatch between the semiconducting material used for the quantum dot itself and for the surrounding barrier – producing strain needed for the spontaneous growth of the dots) or the growth conditions (temperature, deposition rate, subsequent processing such as annealing, etc.).

Although the number of potentially applicable semiconducting materials is huge (consider all elemental semiconductors, but also the wealth of known binary, ternary, … compounds), the nanostructures are often grown using those materials for which the present technology allows for sufficient accuracy in growth of pure nanocrystals or in the epitaxial growth of essentially defect-free quasi-two-dimensional layers, or for sufficient precision in the further necessary nanolithography. Among the commonly used elemental semiconductors are Si, Ge, and C (the latter used in form of natural nanoparticles called fullerenes – for 0D, nanotubes – for 1D, or graphene or ultra-thin graphite – for 2D). The most familiar compounds are GaAs (most often matched with AlGaAs, both having nearly the same lattice constant and thus allowing for the essentially strain-free interfaces) or InGaAs (used for the opposite reason – as having quite strong lattice mismatch with GaAs, and thus offering large strain, needed for spontaneous formation of so-called self-assembled dots) – all representing the III-V compounds. Of the II-VI compounds the most common are probably CdTe and ZnSe, although many others are used as well.



Self-assembled quantum dot

2. Fabrication methods

The methods used to obtain quantum dots can be naturally divided into those following either the top-down or bottom-up approach. In the first one, the dot is achieved progressively, in a sequence of steps, from a larger (often more-dimensional) structure, by such means as nano-lithography or application of external potentials. In the second approach, a spontaneous growth of a useful nanostructure is achieved by a natural self-propelled mechanism, of course driven by a tendency of the carefully designed macro-system to minimize its relevant (local) energy. In general, creation of quantum wires or dots, which confine the carriers to a space with at least two of three dimensions limited to the range of the de Broglie wavelength, requires far more advanced technology than fabrication of quasi-2D quantum wells.

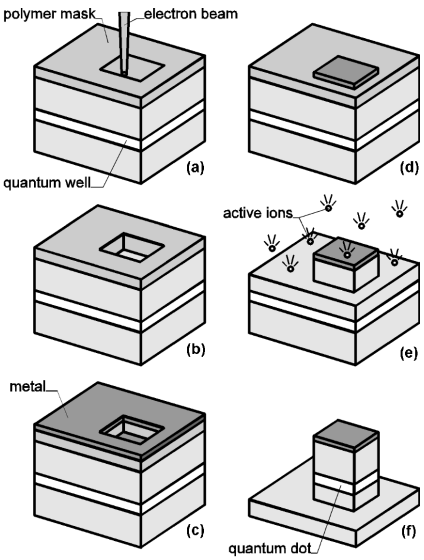
The earliest method of obtaining quantum dots was implemented by Reed et al. [1], who etched them in a structure containing two-dimensional electron gas. The process proceeds in the following steps: The surface of a sample containing one or more quantum wells is covered with a polymer mask, and then partly exposed. The exposed pattern corresponds to the shape of the created nanostructure. Because of the required high resolution, the mask is not exposed to visible light, but to the electron or ion beam (electron/ion beam lithography). At the exposed areas the mask is removed. Later, the entire surface is covered with a thin metal layer. Using a special solution, the polymer film and the protective metal layer are removed, and a clean surface of the sample is obtained, except for the previously exposed areas, where the metal layer remains. Next, by chemically etching the areas not protected by the metal mask, the slim pillars are created, containing the cut-out fragments of quantum wells. In this way, the motion of electrons, which is initially confined in the plane of the quantum well, is further restricted to a small pillar with a diameter on the order of 10-100 nm. A doped base serves as the source of free carriers, which flow into the quantum wells, created above the buffer layers and separated by the barriers. The etching depth drops beneath the interface between the last quantum well and the buffer layer. A metal mask that remains after the etching process may serve as the top electrode. The voltage applied to the electrode controls the number of carriers confined in the dots. The simplicity of producing thin, homogeneous quantum wells makes GaAs the most commonly used material for creating dots by means of etching.

Another method consists in the creation of miniature electrodes over the surface of a quantum well by means of lithographic techniques. The application of an appropriate voltage to the electrodes produces a spatially modulated electric field, which localizes the electrons within a small area. The lateral confinement created in this way shows no edge defects, characteristic of the etched structures. An electric gate can also be created around the etched dot, thus allowing, at least partly, the elimination of edge defects and additional squeezing of electrons.

The process of spreading a thin electrode over the surface of a quantum well may produce either single quantum dots [2,3] or large arrays (matrices) of dots [4,5,6,7]. Modulation of the electric potential, produced by applying voltage to an electrode, can be realized by a previous preparation (using a lithographic technique) of a regular array of islets of nonmetallic material (e.g., of the barrier material) on the surface of the sample. As a result, the distance between the electrode (overlying the surface with the islets) and the plane of the quantum well is modulated, and the electrons are bound in small areas under the prepared islets.

Instead of modulating the distance between the electrode and the well, it is also possible to build a pair of parallel, thin electrodes above the well. The lower electrode can have regularly

placed holes, which is where quantum dots are to be created [8,9]. If a voltage is applied to the pair of electrodes, the result is a change both in the dot size and the depth of the confining potential. The potential depth influences the number of confined electrons. However, when the additional electrode is introduced between the quantum-well layer and the doped layer, the number of electrons and the potential depth can be changed independently.



Etching of a quantum dot

A very advantageous feature of quantum dots whose electrons are confined by the electric field produced by a set of electrodes is their smooth lateral confinement, showing no edge effects. Possibility of controlling certain parameters is also very important. In experimental reports one can find information about these types of dots, which can be created on, for example, GaAs, InSb, or Si

Brunner et al. [10] describe a method for obtaining quantum dots based on a quantum-well material by local heating of a sample with a laser beam. A parent material of a single, 3 nm thick GaAs quantum well was used, and this was prepared using the molecular beam epitaxy method (MBE). It was then placed between a pair of 20 nm thick $Al_{0.35}Ga_{0.65}As$ barriers. The topmost 10 nm thick GaAs cap layer was covered with a 100 nm coating of Si_3N_4, protecting the surface against oxidation or melting by the laser beam. Modulation of the band gap in the quantum well was obtained by local heating on the sample with an argon-laser beam with a power of 5.5 mW. The laser beam was guided along a rectangular contour surrounding an unilluminated area of diameter 300-1000 nm. At a temperature of about 1000 C a rapid interdiffusion of Al and Ga atoms occurred between the well and the barriers, which led to the creation of a local modulation of the material band structure, i.e., to the creation of the potential barrier, which surrounds the unilluminated interior of the rectangle. For larger dimensions of the illuminated rectangle the obtained effective potential that confines the electrons was flat inside the dot (this area will be called A -- pure GaAs), and steep near the

edge (area B – a solution of AlGaAs replaced pure GaAs due to heating). With a decrease of the illuminated rectangle, the area A shrinks. According to the authors, for dimensions near 450 nm the effective potential confining electrons is close to an isotropic parabola. However, it should be mentioned that the details of the electron confining potential in a quantum dot of any type cannot be measured directly (except for the geometric dimensions) and are alternatively obtained through the interpretation of various indirect effects, related to the electronic structure of the object.

It is also possible to create quantum dots in the form of semiconductor microcrystals immersed in glass dielectric matrices. In the first experiment based on that idea, carried out by Ekimov et al. [11], silicate glass with about 1% addition of the semiconducting phase (CdS, CuCl, CdSe, CuBr) was heated for several hours at a temperature of several hundred degrees Celsius, which led to the formation of appropriate microcrystals of almost equal sizes. Knowledge of the dependence of the average crystal radius on the temperature and heating time allowed for controlling their size. The radii of dots measured in different samples varied in the range 1.2-38 nm. As a dielectric matrix, the alkaline chlorides can be used instead of glass. The heating of such a matrix with the addition of copper leads to the formation of CuCl microcrystals.

Quantum dots can also be created through the selective growth of a semiconducting compound with a narrower band gap (e.g., GaAs) on the surface of another compound with a wider band gap (e.g., AlGaAs) [12]. The restriction of growth to chosen areas is obtained by covering the surface of the sample with a mask (SiO_2) and etching on it miniature triangles. On the surface that is not covered with the mask the growth is then carried out with the metal-organic chemical vapor deposition method (MOCVD), at a temperature of 700-800 C. The crystals that are created have the shape of tetrahedral pyramids, and hence when the first crystallized layers are the layers of the substrate compound (AlGaAs) and only the top of the pyramid is created of GaAs, it is possible to obtain a dot of effective size below 100 nm.

A different variant of the method of selective growth is described by Lebens et al. [13]. Onto the 2 μm thick $Al_{0.38}Ga_{0.62}As$ substrate a layer of 10 nm thick GaAs was deposited and covered with a 20 nm Si_3N_4 mask. The mask was later illuminated in chosen areas with the electron beam and removed through the plasma etching. In the growth process, which is carried out with the metal-organic vapor phase epitaxy method (MOVPE), GaAs sedimented only outside the areas covered with the mask. The thickness of the crystallizing GaAs layers was determined to be 100 nm. After covering the newly created structure with a layer of

$Al_{0.2}Ga_{0.8}As$, quantum dots of diameter 70-300 nm and quantum wires of width 90-300 nm and length ~0.1 mm were obtained.

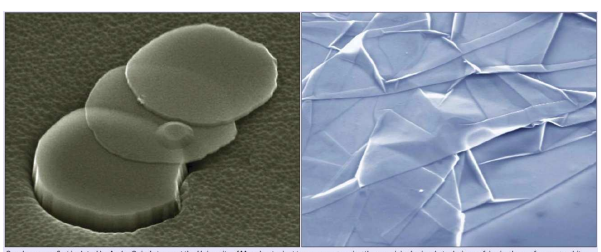
Petroff and DenBaars [14] describe another method for the self-crystallization of quantum dots that does not require the creation of a mask. When the lattice constants of the substrate and the crystallized material differ considerably (7% in the case of GaAs and InAs, the most commonly used pair of compounds), only the first deposited monolayers crystallize in the form of epitaxial, strained layers with the lattice constant equal to that of the substrate. When the critical thickness is exceeded, a significant strain occurring in the layer leads to the break-down of such an ordered structure and to the spontaneous creation of randomly distributed islets of regular shape and similar sizes. The shape and average size of islets depend mainly on factors such as the strain intensity in the layer as related to the misfit of lattice constants, the temperature at which the growth occurs, and the growth rate. The phase transition from

the epitaxial structure to the random arrangement of islets is called the Stranski-Krastanow transition. For the maximum misfit of lattice constants (InAs and GaAs), the transition occurs at the 1.8 monolayer deposition. When the process of crystallization is terminated shortly after reaching the phase transition, the islets evolve to the state of quasi-equilibrium, in which they assume the shape of pyramids or flat, circular lenses, formed on a thin layer of InGaAs (the wetting layer). When the inhomogeneous InGaAs layer is eventually covered with another GaAs layer, a structure of a quantum well with significantly increased thickness in very small areas is obtained.

Raymond et al. [15] reported growth of self-assembled dots in the shape of lenses with ~36 nm diameter and ~4.4 nm height (with fluctuations of 5-10%). Marzin et al. [16] obtained dots in the shape of regular pyramids with a square base of side ~24 nm and height ~2.8 nm (with fluctuations of ~15%), and distance between neighboring dots ~55 nm.

The quantum dots formed in the Stranski-Krastanow phase transition are called self-organized or self-assembled dots (SAD). The small sizes of the self-assembled quantum dots (diameters in the range of 30 nm or even smaller), homogeneity of their shapes and sizes in a macroscopic sample, perfect crystal structure (without edge defects), and the fairly convenient growth process, without the necessity of the precise deposition of electrodes or etching – are among their greatest advantages. Thus, from the very beginning there was great (and now fulfilled) hope regarding their future application in electronics and opto-electronics.

Most recently, a new generation of quantum dots are expected to be soon obtained from graphene – a novel quasi-two-dimensional material in form of atomically thin single honeycomb layer of carbon (for its pioneering experimental studies, Geim and Novoselov jointly received the 2010 Nobel prize in physics). Relativistic nature of carriers (electrons and holes) in graphene makes their lateral confinement more difficult than in conventional semiconductors. Namely, the so-called Klein tunneling paradox forbids confinement by means of a simple potential barrier, such as those produced by the electrodes. However, confinement of initially 2D electrons to quasi-1D nano-strips or quasi-0D nano-islands remains possible by physically cutting such structures out of en extended sheet of graphene. Among the outstanding issues is the stability of the edge of such terminated graphene structures (e.g., the possibility of edge passivation with hydrogen). Intense theoretical work in this direction, driven largely by the unique electronic and mechanical properties of (extended) graphene but also by its natural cleanliness and reduced dimensionality, promises fast progress also in fabrication technology.



Graphene was first isolated by Andre Geim's team at the University of Manchester just two years ago using the surprisingly simple technique of ripping layers from a graphite surface using adhesive tape. By repeatedly peeling away thinner layers (left), single-atom-thick sheets were obtained (right), as shown in these scanning electron micrographs.

Graphene

3. Electronic structure

Electronic structure of quantum nanostructures is generally well described within the effective mass theory, with effective band parameters (energy gaps, band offsets at the interfaces, effective masses) inherited – to a good approximation – from the bulk host semiconductor compounds. If at all needed, modification to these parameters is largely the consequence of strain (which necessarily occurs near any interface separating materials of different lattice constants) or non-uniform composition of the material from which the heterostructure is formed (arising due to interdifussion of atoms across the interface during or after the growth process). Another effect, most significant in small structures, is the leakage of the wave functions of confined carriers into the barriers, leading to the averaging of band parameters of the nanostructure material and of the surrounding barrier material. It need be emphasized that even in the relatively small nanostructures, their finite spatial extent does not invalidate the effective mass description (even though, strictly speaking, this theory requires periodic crystal structure and potentials slowly varying on lengt hscale of the inter-atomic separation).

Characteristic energy and length scales for bulk semiconductors are defined by the effective mass m^* (isotropic for the electrons – for the conduction band involves the atomic s-orbitals; anisotropic for the various types of holes – for the valence band emerges from the p-orbitals), and the effective dielectric constant ε. These parameters enter into the convenient scaling parameters: effective Bohr radius

$$a_B^* = \frac{\varepsilon}{m^*} a_B$$

where $a_B \approx 0.053$ nm is the Bohr radius (in vacuum), and effective Rydberg

$$Ry^* = \frac{m^*}{\varepsilon^2} Ry$$

where $Ry \approx 13.6$ eV is the Rydberg (in vacuum). The large values of the dielectric constant ad well as typically low effective masses (at least for the electron) in the most commonly used semiconductor compounds cause significant enhancement of this effective length scale and significant reduction of this effective energy scale – compared to those relevant for vacuum (i.e., to the hydrogen atom). For example, for bulk GaAs, ε=12.5 (this is the so-called *static* dielectric constant; in general ε depends on frequency) and m^*=0.067 (for the conduction band). This leads to

$$a_B^* \approx 187\, a_B \approx 10 \text{ nm}$$

and

$$Ry^* \approx \frac{1}{2330} Ry \approx 5.8 \text{ meV}$$

.

Needless to say, using effective units removes physical constants from the effective-mass Schrödinger equation, which simplifies to the form:

$$\left[-\nabla^2 + V(\boldsymbol{r})\right]\phi(\boldsymbol{r}) = E\phi(\boldsymbol{r})$$

.

One of the important consequences of the reduced dimensionality of semiconductor nanostructures is a qualitatively different form of the density of states, compared to the bulk

material. Density of states ($\rho = dN/dE$) is defined as the number of available single-particle states per unit energy range (or the inverse of the average energy spacing), taken per unit volume of space. In a three-dimensional medium, the number of electron states up to a wave vector k is

$$N = 2\left(\frac{4}{3}\pi k^3\right)\left(\frac{2\pi}{L}\right)^{-3} L^{-3} = \frac{8\pi}{3}\left(\frac{k}{2\pi}\right)^3$$

Differentiating this with respect to energy gives

$$\frac{dN}{dE} = \frac{dN}{dk}\left(\frac{dE}{dk}\right)^{-1}$$

where

$$\frac{dN}{dk} = \frac{8\pi}{3}\frac{3k^2}{(2\pi)^3} = 4\left(\frac{k}{2\pi}\right)^2$$

Assuming parabolic dispersion (i.e., an isotropic and constant effective mass m*)

$$E = \frac{\hbar^2}{2m^*}k^2$$

we get

$$\frac{dE}{dk} = \frac{\hbar^2}{m^*}k$$

and, finally,

$$\frac{dN}{dE} = 4\left(\frac{k}{2\pi}\right)^2\left(\frac{\hbar^2 k}{m^*}\right)^{-1} = \frac{4m^*}{(2\pi\hbar)^2}k = \frac{4m^*}{(2\pi\hbar)^2}\left(\frac{2m^*E}{\hbar^2}\right)^{1/2} = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}E^{1/2},$$

which shows the well-known square-root (increasing) dependence on energy.

In less than three dimensions the derivation proceeds similarly, but the expression for $N(k)$ is different. Specifically, in two dimensions

$$N = 2\left(\pi k^2\right)\left(\frac{2\pi}{L}\right)^{-2} L^{-2} = 2\pi\left(\frac{k}{2\pi}\right)^2$$

which yields

$$\frac{dN}{dk} = \frac{k}{\pi}$$

and

$$\frac{dN}{dE} = \frac{k}{\pi}\left(\frac{\hbar^2 k}{m^*}\right)^{-1} = \frac{m^*}{\pi\hbar^2}$$

which is independent of k or E. For discrete bands, corresponding to the quantized motion in the perpendicular direction, this becomes:

$$\frac{dN}{dE} = \frac{m^*}{\pi\hbar^2}\sum_i \Theta\left(E - E_i\right)$$

.

In one dimension

$$N = 2(2k)\left(\frac{2\pi}{L}\right)^{-1} L^{-1} = \frac{2k}{\pi}$$

leading to

$$\frac{dN}{dk} = \frac{2}{\pi}$$

and

$$\frac{dN}{dE} = \frac{2}{\pi}\left(\frac{\hbar^2 k}{m^*}\right)^{-1} = \frac{2m^*}{\pi\hbar^2}k^{-1} = \frac{1}{\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2} E^{-1/2}$$

which shows an inverse-square-root divergence. For the quantized motion in the two perpendicular directions, this becomes

$$\frac{dN}{dE} = \frac{1}{\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\sum_i E^{-1/2}\Theta\left(E - E_i\right)$$

.

Finally, in fully spatially quantized, zero-dimensional systems (where the motion in any direction is not characterized by a continuous wave vector *k*), one finds

$$\frac{dN}{dE} \propto \sum_i \delta\left(E - E_i\right)$$

,

that is, the density of states consisting of a series of discrete peaks at the energies corresponding to the quantized levels.

The calculation of single-particle energy levels in particular nanostructures usually involves identification of the mechanism of confinement, definition of the adequate modeling potential, and solution of the corresponding stationary Schrodinger equation. Depending on a particular confinement, its potential can be calculated from first principles or modeled in a convenient phenomenological form, with the physically intuitive parameters treated as free parameters or matched to the available experimental information. Obviously, the problem can often be greatly simplified by taking advantage of relevant symmetries – both simple geometric (e.g., following from the rotational or reflective symmetry of the physical structure) or more subtle, dynamic (e.g., consisting of effective decoupling of degrees of freedom corresponding to the incommensurate characteristic excitation energy scales; another example being the harmonic confinement which results in the degeneracy of energy levels and emergence of the shell structure in the spectrum). The accuracy in modeling the single-particle energy spectrum of the nanostructure depends on its purpose; qualitative features (such as symmetries) are most important in the search of new phenomena (e.g., arising from interactions in many-electron systems); on the other hand, accuracy may be essential in designing real devices.

4. Comparison with atomic and macroscopic systems

The defining properties which contrast nanostructures from larger, macroscopic systems are the quantization of energy levels and of the number of confined carriers. Let us stress that the quantization becomes essential when its characteristic energy (dependent not only on size, but also on effective mass) exceeds other relevant energy scales such as the average phonon energy (controlled by the lattice temperature).

Compared to smaller systems, such as natural atoms or molecules, nanostructures show more pronounced interaction and correlation effects, owing to a different ratio of the characteristic single-particle (quantization) energy to the characteristic interaction (correlation) energy. However, it is the size in the effective units that matters, as can be readily understood by estimating crudely the single-particle energy E from a particle-in-a-box model

$$E = \frac{\hbar^2}{2m^*}\left(\frac{\pi}{L}\right)^2$$

and taking the interaction energy scale U as corresponding to the length scale of the system

$$U = \frac{e^2}{\varepsilon L}.$$

Indeed, the correlation-to-quantization energy ratio can be expressed as

$$\frac{U}{E} = \frac{e^2}{\varepsilon L} \Big/ \frac{\hbar^2}{2m^*}\left(\frac{\pi}{L}\right)^2 \propto \frac{m^*}{\varepsilon} L \propto \frac{L}{a_B^*},$$

meaning that similar correlation effects as in single natural atoms will occur in semiconductor nanostructures extending many atoms in length.

Thus, Hund rules predicting few-electron correlation effects in degenerate atomic shells are also valid in small symmetric dots, but inter-shell correlations begin to play the dominant role in larger structures.

A rather obvious difference between artificial and natural atoms or molecules is a far greater – in the former case – available variation in their size, shape (including geometrical symmetry), and dynamical properties of the confinement (Coulomb, hard-wall, nearly harmonic – leading to the presence or absence of a particular shell structure and to the emergence or removal of the particular selection rules for the relevant internal transitions). A practically important tool in tuning electronic (or magnetic) properties of quantum dots is by means of nanodesign of their interior – for example by placing an appropriate number of magnetic atoms inside the dot, which on one hand can be addressed from outside by external magnetic field and on the other – couple magnetically to the confined electrons or holes.

Another useful distinction lies in the separation of electron and hole bands (in semiconductor structures – convenient from the point of view of possible resonant excitation, often by means of simple photo-absorption in the range of wave lengths close to the visible light). Also the electron-hole symmetry is not exact (thus, conveniently tunable) in the semiconductors, in contrast to the fixed electron-positron symmetry in vacuum (i.e., in natural atoms). Thus, semiconductor structures can be rather easily populated with positive and negative carriers by

means of photo-excitation (practically, illumination), and the photo-carriers can be made live sufficiently long to allow their manipulation, for example by means of electric potentials.

Finally, natural atoms can be thought of as essentially isolated when considered in vacuum. In contrast, semiconductor nanostructures are always embedded in (and thus coupled to) the surrounding medium which is a thermodynamic environment characterized by its own length and energy scales. And so, phonon effects cannot be ever completely eliminated in quantum dots, often undermining a simple purely quantum mechanical picture of a controlled number of electrons confined by an effective and static potential.

5. Role of dimensionality; extended versus fully confined systems

Reduced dimensionality (compared to bulk, or three dimensions) allows novel realizations of quantum mechanics in the sense of quantum statistics. Namely, the only statistics allowed in three dimensions are those of Bose and Fermi, corresponding to the phase of a many-body wave-function merely acquiring a phase of +1 (i.e., no phase) or -1 upon an exchange of a pair of particles. This, of course, is the "topological" contribution to the phase change, resulting from the fact that the exchange took place, and not from the way it did. The wave-function acquires also a trivial phase $\exp(-\omega t)$ associated with the time passage, unrelated to a possible exchange during this time. It also may acquire a so-called geometric (or Berry) phase, connected to the paths swept by the individual particles, also unrelated to whether the particles exchange or not. Thus, in three dimensions, the "topological" effect of an exchange of a pair of particles on the many-body wave function can only be either null (Bosons) or a phase factor of -1 (Fermions).

The proof is straightforward and relies on the topological equivalence of (the paths of) a double exchange among the pair of particles and a closed loop of one particle around the other, which *exclusively in three dimensions* is also topologically equivalent to identity (i.e., null operation, or no exchange at all). Clearly, the assumption of the latter equivalence is no longer valid in two dimensions, where a loop around a particle cannot be continuously removed, and thus reduced to a null operation. Hence, the topological effect of a double exchange may emerge, resulting in a wider variety of statistics, besides the Bosons or Fermions.

The precise mathematical description of the adiabatic exchanges among the identical particles involved the concept of a braid group. Each "braid" – an element of this group – represents a topologically unique spatial transformation of a certain number of particles N from their initial positions $R_1, R_2, ..., R_N$ to the same set of positions (but, of course, not all the particles need end up in their individual initial positions). Braid group is distinct (larger) from the permutation group. Indeed, a braid represents not only the permutation but also the topological class of the path along which the exchange took place (e.g., which particle made a loop around which).

The fact that Fermions and Bosons are the only quantum statistics allowed in three dimensions can be expressed as that the only relevant irreducible representations of the braid group are one-dimensional and, moreover, they only include two elements: 1 and -1, such that each element squared results in an identity.

In contrast, in two spatial dimensions, more complex representations also become relevant. Among the one-dimensional representations the lack of requirement that each element

squared must result in an identity allows for the braids represented by an arbitrary exchange phase θ, so that the factor multiplying the many-body wave-function in result of an exchange can be anything, $\exp(i\theta)$, not only -1 or 1. Such hypothetical particles are called Anyons. They were first postulated by Leinaas and Myrheim [17], later studied (and actually named) by Wilczek [18], and currently believed to occur in the so-called fractional quantum Hall systems [19]. Specifically, the elementary carrier of an electric charge in the so-called Laughlin liquid [20] (an incompressible quantum liquid phase of electrons formed in two dimensions and in a high perpendicular magnetic field) has been convincingly demonstrated [21] to carry a fractional charge e/3, corresponding to a fractional exchange phase of $\theta=\pi/3$.

However, even more exotic quantum statistics are also possible, corresponding to the multi-dimensional representations of the braid group. In such case, more than a single quantum state must correspond to the same spatial configuration (R_1, R_2, ..., R_N) of a set of quantum particles. These different quantum states need obviously be distinguished by "topological" rather than orbital quantum numbers. Let's denote their number by g. Thus, an arbitrary state with the spatial configuration (R_1, R_2, ..., R_N) is a linear combination of those g basis states. It can be conveniently represented by a g-dimensional vector in this basis. Now, a braid, which stands for a particular adiabatic position exchange among the N particles, must describe the fact that although such an exchange must bring the N particles back to their original spatial configuration, it does not have to restore the original quantum state (represented by the g-dimensional vector). Thus, the topological effect of the process of braiding consists of a rotation of the vector state in the corresponding g-dimensional Hilbert space. And hence, such a braid must be represented by a g-dimensional square matrix, rather than by a simple phase factor $\exp(i\theta)$. Since multiplication of the matrices need not be Abelian, neither need be the actual topological effect of the sequence of exchanges (braids) among the particles. Thus, the various possible corresponding quantum statistics are generally called non-Abelian, and the hypothetical quantum particles obeying such statistics – non-Abelions.

Whether the non-Abelian quantum particles occur in nature remains an open fundamental question, and an object of intensive current studies (both theoretical and experimental). The most plausible candidates are the quasiparticle excitations of another (non-Laughlin) electron liquid formed in the fractional quantum Hall regime. The so-called Pfaffian wave-function proposed by Moore and Read [22], whose excitations have been demonstrated to be non-Abelian, appears (mostly from the numerical simulations) to capture the essence of the real many-electron state in the corresponding conditions. By capturing the essential behaviour we mean an adiabatic connection between the model wave-function and the actual physical state, and thus a correct description of the qualitative/topological features, such as the quantum statistics.

Thus, spatial confinement of electrons from three down to two dimensions opens the door to topological many-body effects associated with braiding. This also remains true in the lower-dimensional nanostructures, as long as the confinement in one direction (typically that of growth) is much stronger than in the other (lateral) directions. In such case, the lateral confinement may lead to complete spatial quantization of the electrons, but without a complete destruction of the topological properties of the initially two-dimensional state.

Finally, strong confinement to only one direction removes the possibility of adiabatic position exchange altogether, again resulting in a completely new topological characteristics.

II. Overview of computational methods

1. Many-body states of non-interacting particles

Let us consider a system of N identical, non-interacting particles, the dynamics of each of them being governed by a single-particle Hamiltonian h(x), expressed through a composite variable x (including spatial and spin coordinates). The single-particle eigenstates are then obtained from the Schrödinger equation

$$h(x)\,u_k(x) = \varepsilon_k\,u_k(x)$$

In the absence of interactions, the many-body Hamiltonian is simply an uncoupled sum of single-particle Hamiltonians

$$H = \sum_{i=1}^{N} h(x_i)$$

The solution to the corresponding many-body Schrödinger equation has the form of a product of the single particle states (called spin-orbitals)

$$\psi = \prod_{i=1}^{N} u_{k_i}(x_i)$$

The energy of this state is simply the sum of the corresponding single particle energies

$$E = \sum_{i=1}^{N} \varepsilon_{k_i}$$

For the Fermions, which obey the Pauli exclusion principle (forbidding any pair to occupy the same spin-orbital), the ground state can be constructed from the sequence of the lowest N spin orbitals

$$\psi_{GS} = \prod_{k=1}^{N} u_k(x_k)$$

with the corresponding ground state energy

$$E_{GS} = \sum_{k=1}^{N} \varepsilon_k$$

However, Fermions not only obey the Pauli exclusion principle, but do so as a consequence of the requirement that a pair exchange results in a minus sign acquired by the wave-function. Hence, the many-Fermion wave-function must be completely anti-symmetric with respect to all pair exchanges. The above expressions for ψ clearly are not. The anti-symmetrization can readily be applied to the above expressions, to result in the so-called Dirac-Slater determinant form of the many-Fermion wave-function, i.e., a determinant involving N Fermions distributed in all possible (N!) ways among the N select spin-orbitals

$$\psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} u_{k_1}(x_1) & u_{k_1}(x_2) & \cdots & u_{k_1}(x_N) \\ u_{k_2}(x_1) & u_{k_2}(x_2) & \cdots & u_{k_2}(x_N) \\ \vdots & \vdots & \ddots & \vdots \\ u_{k_N}(x_1) & u_{k_N}(x_2) & \cdots & u_{k_N}(x_N) \end{vmatrix}$$

For the ground state, this becomes

$$\psi_{GS} = \frac{1}{\sqrt{N!}} \begin{vmatrix} u_1(x_1) & u_1(x_2) & \cdots & u_1(x_N) \\ u_2(x_1) & u_2(x_2) & \cdots & u_2(x_N) \\ \vdots & \vdots & \ddots & \vdots \\ u_N(x_1) & u_N(x_2) & \cdots & u_N(x_N) \end{vmatrix}$$

to reflect the occupation of the lowest N spin-orbitals.

For Bosons, whose wave function need be symmetric with respect to pair exchanges, the expression for ψ is analogous to the above, except that the determinant is replaced by the permutant, and the normalization prefactor is adjusted for the case when more than one Boson occupies the same spin-orbital.

In particular, the many-Boson ground state is a condensate in the lowest spin orbital, obtained by setting $k_1 = k_2 = \ldots = k_N = 1$.

Importantly, the symmetrization or anti-symmetrization only affects the form od the wave-function, while the energy remains the sum of the single-particle energies, insensitive to the statistics.

2. Hartree approximation

In the presence of pair interaction among the particles, defined by its potential $v(x_1,x_2)$, the many-body Hamiltonian takes the form

$$H = \sum_{i=1}^{N} h(x_i) + \frac{1}{2} \sum_{\substack{i,j=1 \\ (i \neq j)}}^{N} v(x_i, x_j)$$

.

Here, the factor (1/2) adjusts for the double counting of each pair (i,j) and (j,i). For the interaction dependent only on the relative position (i.e., on distance) $r_{12}=|\mathbf{r}_1-\mathbf{r}_2|$, such as Coulomb interaction among the electrons

$$v(r_{ij}) = \frac{e^2}{r_{ij}}$$

(here, the dielectric constant $4\pi\varepsilon\varepsilon_0$ is set to unity) this simplifies to

$$H = \sum_{i=1}^{N} h(x_i) + \frac{1}{2} \sum_{\substack{i,j=1 \\ (i \neq j)}}^{N} v(r_{ij})$$

.

The above Hamiltonian appears unstable (towards either collapse or expansion, depending on the sign on the interaction), but in real many-electron systems the stability is secured by the

presence of a charge neutralizing background of the nuclei. Specifically, Coulomb attraction among the electrons and ions is hidden in the single particle Hamiltonian

$$h(x) = \frac{p^2}{2m} + \sum_{\alpha=1}^{N} V(r_\alpha)$$

where α labels the ions, and r_α is the distance between a given electron and the α^{th} ion.

At a fixed ion density (which also determines the electron density), the remaining term in the Hamiltonian which accounts for the ion-ion interaction is a constant, and it can be omitted. However, it need be remembered that the above many-electron Hamiltonian H must be considered at a fixed electron density.

The presence of an interaction term makes it impossible to factorize the Hamiltonian and obtain a general and analytic solution. In the approximate approach due to Douglas Hartree, it is assumed that each electron moves in an average Coulomb field of the other electrons. This leads to the following single-particle Schrödinger equation

$$\left[h(x_k) + \sum_{k' \neq k} \int dx_{k'}\ v(r_{kk'}) \left| u_{k'}(x_{k'}) \right|^2 \right] u_k(x_k) = \varepsilon_k\ u_k(x_k)$$

.

This so-called Hartree equation can be rewritten as follows by rearranging the involved sum (adding and subtracting the k=k' term)

$$\left[h(x) + V_H(x) - \Sigma_k^{SI}(x) \right] u_k(x) = \varepsilon_k\ u_k(x)$$

.

Here,

$$V_H(x) = \int dx'\ v(r,r') n(x')$$

is the Hartree potential expressed through the total electron density

$$n(x) = \sum_{k'} n_{k'} \left| u_{k'}(x_{k'}) \right|^2$$

which in turn involves the occupation numbers $n_{k'}$=1 or 0, depending on whether the $(k')^{th}$ space orbital is occupied or not.

The other term

$$\Sigma_k^{SI}(x) = \int dx'\ v(r,r') \left| u_k(x') \right|^2$$

is called the self-interaction correction, as it removes the (unphysical) k=k' term added in the definition of V_H.

The Hartree approximation provides quite an accurate description of the electronic structure of isolated atoms, and also allow the estimates of atomic electron densities n(x) which agree fairly well with those obtained experimentally from the X-ray scattering.

The formal justification of the method comes from the observation that it is equivalent to the search for the set of (so-called Hartree) spin-orbitals $u_k(x)$ such that the product many-electron wave function

$$\psi_{GS} = \prod_{k=1}^{N} u_k(x_k)$$

minimizes the average energy (expectation value of the Hamiltonian). Let us stress that this is an optimal set of spin-orbitals for the lowest-energy many-body wave function which is not anti-symmetrized.

The search proceeds using the variational principle. The relevant variational equation is

$$\langle \delta\psi | E - H | \psi \rangle = 0$$

where $\delta\psi$ is an infinitesimal variation of ψ, obtained by varying separately each of the involved u_k by δu_k.

The Hartree method replaces the N-electron problem by an effective single-electron problem, but at a cost of using in the Schrödinger equation the unknown quantity

$$n(x') - |u_k(x')|^2 .$$

The equations must hence be solved iteratively – beginning with a starting guess for the set of spin-orbitals u_k, which are then used to calculate the Hartree and self-interaction potentials, from which a better approximation to the optimal set of spin-orbitals u_k can be calculated by solving the single-particle Schrödinger equation. The procedure must then be repeated until reaching self-consistency (i.e., the convergence of the energies ε_k and spin orbitals u_k).

Douglas Rayner Hartree (1897-1958)

3. Hartree-Fock approximation

This method, introduced by Douglas Hartree and Vladimir Fock, is an improvement over the Hartree approximation in that it accounts for the requirement of that the many-Fermion (e.g., many-electron) wave function must be anti-symmetric. Being somewhat less intuitive, it is best explained in the variational language, as follows.

Like in the Hartree method, the goal of the Hartree-Fock method is to find an optimal set of spin-orbitals $u_k(x)$, such that the average total energy is minimized with respect to the anti-symmetrized products of those spin-orbitals. Total energy means here the total single-particle and interaction energy, and the average total energy means the expectation value of the many-electron Hamiltonian H.

$$H = \sum_{i=1}^{N} h(x_i) + \frac{1}{2} \sum_{\substack{i,j=1 \\ (i \neq j)}}^{N} v(r_{ij})$$

On the other hand, the anti-symmetrized product of spin-orbitals $u_k(x)$ is the Dirac-Slater determinant introduced earlier

$$\psi_{GS} = \frac{1}{\sqrt{N!}} \begin{vmatrix} u_1(x_1) & u_1(x_2) & \cdots & u_1(x_N) \\ u_2(x_1) & u_2(x_2) & \cdots & u_2(x_N) \\ \vdots & \vdots & \ddots & \vdots \\ u_N(x_1) & u_N(x_2) & \cdots & u_N(x_N) \end{vmatrix}.$$

The average value of H in such a state is compactly expressed in the "bra-ket" notation, where the "ket" $|k\rangle$ stands for the single particle state corresponding to the wave-function $u_k(x)$, and the "bra" $\langle k|$ denotes the complex conjugate $u^*_k(x)$. For example, the scalar product of states k and k' is written as

$$\int dx\, u_k^*(x)\, u_{k'}(x) = \langle k | k' \rangle$$

and the matrix element of a single-particle operator h becomes

$$\int dx\, u_k^*(x)\, h(x)\, u_{k'}(x) = \langle k | h | k' \rangle .$$

Analogously, pair states are denoted as follows

$$u_k(x)\, u_{k'}(x') \sim |k, k'\rangle$$

$$u_k^*(x)\, u_{k'}^*(x') \sim \langle k, k' |,$$

so that the two-body interaction matrix element becomes

$$\int dx \int dx'\, u_k^*(x)\, u_{k'}^*(x')\, v(x,x')\, u_l(x)\, u_{l'}(x') = \langle k, k' | v | l, l' \rangle .$$

Using this notation, the average total energy (expectation value of H) in the many-body state spanned by the set of spin-orbitals $u_k(x)$ can be written as

$$E \equiv \langle H \rangle = \sum_k n_k \langle k | h | k \rangle + \frac{1}{2} \sum_{k,l} n_k n_l \left[\langle k, l | h | k, l \rangle - \langle k, l | h | l, k \rangle \right],$$

where n_k=1 or 0 is the occupation of the spin-orbital u_k, as defined earlier. The first interaction term inside the square bracket is called the direct term, and it gives the average interaction energy of a pair of distinguishable particles occupying spin-orbitals k and l (distinguishable – therefore the pair wave-function is a simple product, without anti-symmetrization). The latter term is called an exchange term, as it gives the amplitude of an interchange of a pair of identical particles in spin-orbitals k and l, caused by their mutual interaction. This term emerges for a pair of identical Fermions, as a result of their wave-function being antisymmetric

$$\psi_{k,k'}(x,x') = \frac{1}{\sqrt{2}}\left[u_k(x)\,u_{k'}(x') - u_k(x')\,u_{k'}(x)\right]$$

.

The presence of an exchange term in the expression for E is the only difference between the Hartree and Hartree-Fock methods.

The variational procedure proceeds as follows. The goal is to minimize E with respect to the variation of $u_k(x)$ by an arbitrary $\delta u_k(x)$, with an additional constraint that the set of spin-orbitals remains ortho-normal

$$\langle k | k' \rangle = \delta_{k,k'}$$

.

This is achieved by using the method of Lagrange multipliers ($\lambda_{k,k'}$), giving the following relation to be satisfied for an arbitrary first-order change $\delta u^*_k(x)$

$$\delta E - \sum_{k,k'} \lambda_{k,k'} \int dx\; \delta u^*_k(x)\, u_{k'}(x) = 0$$

.

This leads to the following set of single-particle equations

$$\left[h(x) + \sum_l n_l \int dx'\, v(r,r') \left| u_l(x') \right|^2 \right] u_k(x)$$

$$- \left[\sum_l n_l \int dx'\, v(r,r')\, u^*_l(x')\, u_k(x') \right] u_l(x) = \sum_l \lambda_{kl}\, u_l(x)$$

in which the first of the two integrals is the Hartree potential V_H, and the second one is called the exchange term. These equations can be simplified by applying a unitary transformation to ensure a diagonal form

$$\lambda_{kl} = \varepsilon_k \delta_{kl}$$

.

The simplified equations read

$$\left[h(x) + V_H(x) \right] u_k(x) + \int dx'\, \Sigma_X(x,x')\, u_k(x') = \varepsilon_k u_k(x)$$

where $\Sigma_x(x,x')$ is the (non-local) exchange potential defined as

$$\Sigma_X(x,x') = -v(r,r') \sum_l n_l\, u_l(x')\, u^*_l(x')$$

.

The l=k term in $\Sigma_x(x,x')$ is the self-interaction term known from the Hartree approximation. The remaining k≠l exchange terms make the difference between the Hartree and Hartree-Fock approximations.


Vladimir Aleksandrovich Fock

4. Correlations and entanglement

By construction of the many-body states in the form of the products of single-particle spin-orbitals (whether anti-symmetrized or not), both Hartree and Hartree-Fock approximations exclude the effect of correlation. Such correlation (among the motion of different particles) would manifest itself by the dependence of the many-body wave-function on the relative positions (distances), evidently absent in the product expression.

Let us illustrate this on a simple example for two particles. In the product pair state, whether anti-symmetrized

$$\psi_{k,k'}(x,x') = \frac{1}{\sqrt{2}}\left[u_k(x)\,u_{k'}(x') - u_k(x')\,u_{k'}(x)\right]$$

or not

$$\psi_{k,k'}(x,x') = u_k(x)\,u_{k'}(x')\,,$$

one particle is in state u_k regardless of the other particle, and the other is in state $u_{k'}$ regardless of the first. This is precisely the lack of (their) correlation. Let us denote this state (anti-symmetrized or not) as

$$|k,k'\rangle$$

In contrast, in the following state

$$|k,k'\rangle + |l,l'\rangle + \ldots$$

one particle is in the state u_k only if the other is in the state $u_{k'}$ (and vice versa), one is in the state u_l only if the other is in $u_{l'}$ (and vice versa), etc. This is a perfectly correlated state, in a sense that the state of one particle completely determines the state of the other. Another term commonly used in this context is entanglement.

Note also that the correlation does not hide in the fact that there are more than one term forming a linear combination. As a counter-example, let us consider the following state

$$|k,k'\rangle + |l,l'\rangle + |k,l'\rangle + |l,k'\rangle .$$

It is perfectly uncorrelated in a sense that the state of one particle (e.g., u_k versus u_l) has no influence on the state of the other particle ($u_{k'}$ versus $u_{l'}$). In other words, detection of one particle in either of the two allowed states (e.g., u_k or u_l) provides no information about the state of the other particle ($u_{k'}$ or $u_{l'}$).

Rather obviously, states of intermediate degree of correlation (entanglement) are also possible in which state of one particle determines the states of the other one only to some extent.

5. Correlation hole

Let us consider the effect of electron correlation in an extended system (where an exact numerical treatment is not applicable). The idealized system to consider is the free electron gas. Relevant to simple metals, this model has nuclear charges evenly smeared out in space and forming a uniform positive background, compensating negative charge of the electrons. Hence, in the Hartree approximation, each electron feels no force, or a constant (zero) potential in the whole space. The solutions of the single-particle Hartree equations are

$$u_k(r) = V^{-1/2} e^{ikr}$$

$$\varepsilon_k = \frac{\hbar^2 k^2}{2m}$$

where V is the volume and m is the electron mass.

In the ground state, these states are filled up from k=0 to the highest one k= k_F called Fermi momentum, each with two electrons of the opposite spin. Of course, the Fermi momentum depends on density n

$$\frac{4}{3}\pi\left(\frac{k_F}{2\pi}\right)^3 = \frac{1}{2}n$$

(the factor 1/2 on the right hand side due to the spin degeneracy), and the Fermi energy is

$$\varepsilon_F \equiv \varepsilon_{k_F} = \frac{\hbar^2 k_F^2}{2m} .$$

A convenient measure of density is the average radius per electron r_s defined as

$$\frac{4}{3}\pi\, r_s^3 = \frac{1}{n} .$$

Using this quantity, one can write

$$k_F = \frac{1}{\alpha r_s}$$

with

$$\alpha = \left(\frac{4}{9\pi}\right)^{1/3} = 0.521$$

.

Let us now observe that the plane waves are the eigenstates of the exchange term in the Hartree-Fock equation

$$\int dr' \, \Sigma_X(r,r') u_k(r') = (\Sigma_X)_{k,k} u_k(r)$$

where

$$(\Sigma_X)_{k,k} = -\frac{4\pi e^2}{V} \sum_{k'} \frac{n_{k'}}{|k-k'|^2}$$

This allows the calculation of an average exchange energy per electron ε_x

$$\varepsilon_x = \frac{1}{N}\frac{1}{2}\sum_{k \le k_F} (\Sigma_X)_{k,k} = -\frac{2\pi e^2}{NV} \sum_{k,k' \le k_F} \frac{n_{k'}}{|k-k'|^2}$$

(the factor 1/2 for the double counting of all pairs). The result is

$$\varepsilon_x = -\frac{3e^2}{4\pi\alpha r_s} = -\frac{0.458}{r_s}$$

.

Let us now introduce the concept of a correlation hole. We begin by rewriting the exchange term of the Hartree-Fock equation as

$$\int dx' \, \Sigma_X(x,x') u_k(x') = -\left[\frac{\int dx' \, v(r,r') \sum_l n_l \, u_l(x) u_l^*(x') u_k(x')}{u_k(x)}\right] u_k(x)$$

The term in the square brackets is the single-particle effective exchange potential $V_{HF}(x)$. It can be interpreted as an electrostatic potential attributed to a fictitious electron density

$$n_{HF}(x,x') = \sum_l n_l \frac{u_l(x) u_l^*(x') u_k(x')}{u_k(x)} = \left(\sum_l n_l \, u_l(x) u_l^*(x')\right) \frac{u_k(x')}{u_k(x)}$$

which is normalized to

$$\int dx' \, n_{HF}(x,x') = 1$$

corresponding to the fact that each electron interacts with N-1 other electrons that can be counted as N electrons contained in the Hartree potential V_H minus the unit charge of the so-called exchange hole. What is the density profile of the exchange hole (around its electron)? If all orbitals were occupied (corresponding to an infinite Fermi energy), i.e., all n_l=1, then

$$\sum_l n_l \, u_l(x) u_l^*(x') = \delta(x-x')$$

and thus also

$$n_{HF}(x,x') = \delta(x-x')$$

which would mean a correlation hole which is localized precisely on an electron. For a Fermi sphere extending up to a finite Fermi wave vector k_F, $n_{HF}(x,x')$ is broadened to a characteristic Fermi width of

$$\lambda_F = \frac{2\pi}{k_F} = 3r_s$$

.

In both Hartree and Hartree-Fock approximations, each electron feels an average Coulomb field from the other electrons. However, the electron-electron repulsion creates a correlation hole around each electron. This hole screens the charge of its electron (when immersed in the sea of other electrons) and in this way lowers the total interaction energy.

Let us write the electron density as

$$n(r) = \langle \psi | \sum_{i=1}^{N} \delta(r - r_i) | \psi \rangle$$

i.e., the probability (per unit volume) to find an electron at position r. Here, ψ is the many (N) electron ground state, and r_i are the electron positions.

Analogously, one can define the pair correlation function as

$$n(r,r') = \langle \psi | \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \delta(r - r_i)\, \delta(r' - r_j) | \psi \rangle$$

i.e., the joint probability to find an electron at position r and another electron at position r'.

The pair correlation function can be used to write the total many-body Coulomb energy as

$$V_{Coul} \equiv \langle \psi | \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{e^2}{r_{ij}} | \psi \rangle = \frac{e^2}{2} \int dr \int dr' \, \frac{n(r,r')}{r_{ij}}$$

.

For completely uncorrelated electrons, by definition,

$$n(r,r') = n(r)\, n(r')$$

.

In general, one can introduce a function g(r,r') and write

$$n(r,r') = n(r)\, n(r') \left[1 + g(r,r') \right]$$

.

From the normalization of

$$\int dr' \, n(r,r') = N - 1$$

one immediately obtains the following sum rule

$$\int dr' \, g(r,r')\, n(r') = -1$$

which, again, means that each electron is screened by an exchange-correlation hole dragged with itself.

To understand the relative contributions of exchange and correlation to this hole let us consider a spin-unpolarized system, with $N\uparrow$ and $N\downarrow$ electrons of the respective spin. Clearly,

each spin-↑ electron interacts with N↑-1 electrons of the same spin and with N↓ electrons of the opposite spin. Thus, separating the density and function g(r,r') into two spin components, one gets

$$\int dr'\, g_{\uparrow\uparrow}(r,r')\, n_{\uparrow}(r') = -1$$

$$\int dr'\, g_{\uparrow\downarrow}(r,r')\, n_{\uparrow}(r') = 0$$

These sum rules can be interpreted as that the "same-spin" exchange hole remains equal to -1 also in the presence of correlations, and that while the local "opposite-spin" correlation hole is also equal to -1, the compensating charge must accumulate on the surface so as to ensure the latter condition.

6. Thomas-Fermi model

Thomas-Fermi model is an approximate quantum mechanical theory of the electronic structure calculations that does not directly involve the many-body wave-functions. Instead it is formulated entirely in terms of the electronic density and as such it can be regarded as a precursor to the modern density functional theory. The Thomas-Fermi model is strictly correct only in the limit of an infinite nuclear charge. When used as an approximation for realistic atomic or molecular systems, it yields poor quantitative predictions, even failing to reproduce some general features of the density such as shell structure in atoms and Friedel oscillations in solids. It has, however, found useful applications through the ability to extract qualitative behavior analytically.

The key assumption is that the electrons are distributed uniformly in phase space with two electrons in every

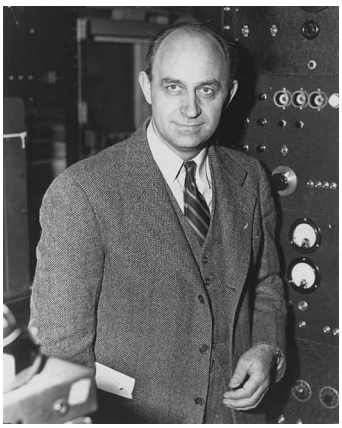
$$(2\pi\hbar)^3$$

of phase volume. In each element of real space volume d^3r, the electrons fill a sphere of momentum space reaching up to the Fermi momentum p_F defined at this given point of real space. The number of electrons inside this Fermi sphere is

$$\frac{\frac{4}{3}\pi p_F^3}{\frac{1}{2}(2\pi\hbar)^3}$$

and the equation of the number of electrons when counted in the real and momentum spaces gives the following key relation between the local density and the local Fermi wave-vector

Llewellyn Thomas (1903-1992)



Enrico Fermi (1901-1954)

$$n(r)=\frac{8\pi}{3}\left(\frac{p_F(r)}{2\pi\hbar}\right)^3=\frac{8\pi}{3}\frac{1}{(2\pi\hbar)^3}p_F^3(r)$$

or, inversely,

$$p_F(r)=2\pi\hbar\left(\frac{3n(r)}{8\pi}\right)^{1/3}=2\pi\hbar\left(\frac{3}{8\pi}\right)^{1/3}n^{1/3}(r)$$

.

The local Fermi energy also depends on the electron density, as follows

$$E_F(r)\equiv\frac{p_F^2(r)}{2m}=\frac{(2\pi\hbar)^2}{2m}\left(\frac{3n(r)}{8\pi}\right)^{2/3}=\frac{(2\pi\hbar)^2}{2m}\left(\frac{3}{8\pi}\right)^{2/3}n^{2/3}(r)$$

.

Substituting this into the classical expression for the kinetic energy

$$T=\int dr^3 n(r)\langle T\rangle_p$$

where an average over a uniformly occupied momentum sphere is

$$\langle T\rangle_p=\frac{\int d^3p\,T(p)}{\int d^3p}=\frac{\int_0^{p_F}4\pi p^2dp\,T(p)}{\int_0^{p_F}4\pi p^2dp}=\frac{\int_0^{p_F}4\pi p^2dp\frac{p^2}{2m}}{\int_0^{p_F}4\pi p^2dp}=\frac{3}{5}\frac{p_F^2}{2m}=\frac{3}{5}E_F$$

,

one obtains the so-called Thomas-Fermi kinetic energy functional

$$T_{TF}[n]=\int dr^3 n(r)\frac{3}{5}E_F(r)=\int dr^3 n(r)\frac{3}{5}E_F(r)=\frac{3}{5}\frac{(2\pi\hbar)^2}{2m}\left(\frac{3}{8\pi}\right)^{2/3}\int dr^3 n^{5/3}(r)$$

.

Or, briefly,

$$T_{TF}[n] = C_F \int d^3r \; n^{5/3}(r)$$

with an adequate constant

$$C_F = \frac{3}{5}\frac{(2\pi\hbar)^2}{2m}\left(\frac{3}{8\pi}\right)^{2/3}.$$

7. Density functional theory

Like the Thomas-Fermi model, density functional theory (DFT) involves the local electronic density n(r) rather than the many-electron wave-functions, and thus expresses all the relevant properties of the system, such as the energy, as functionals of the density function n(r). The DFT theory depends on two Hohenberg-Kohn theorems [23].

The first Hohenberg-Kohn theorem establishes that the ground state properties of a many-electron system are uniquely determined by its electron density (a function of the position). Therefore, at least in principle, it allows to reduce the many-body problem of N electrons with the total number of 3N spatial coordinates to only 3 spatial coordinates, through the use of functionals F[n] of the electron density n(r).

Obviously, the density function n(r) is uniquely determined by the many-body wave-function

$$n(r) = N\int d^3r_2 \int d^3r_3 \ldots \int d^3r_N \, |\psi(r, r_2, r_3, \ldots, r_N)|^2 .$$

According to the first Hohenberg-Kohn theorem, for the ground state this statement can be reversed, to say that the ground state wave function is uniquely determined by the ground state density. In other words, the ground state wave function is the functional of density

$$\psi_{GS} = \psi[n_{GS}(r)].$$

Consequently, any expectation value in the ground state is also a functional of density

$$\langle \psi_{GS} | O | \psi_{GS} \rangle = \langle \psi[n_{GS}(r)] \,|\, O \,|\, \psi[n_{GS}(r)] \rangle = O[n_{GS}(r)].$$

The second Hohenberg-Kohn theorem defines an energy functional for a many-electron system and ascertains that the ground state electron density minimizes this energy functional.

The total energy functional

$$E_{GS} = E[n_{GS}(r)] = \langle \psi[n_{GS}(r)] \,|\, T + V + U \,|\, \psi[n_{GS}(r)] \rangle$$

contains the kinetic energy T, the single particle (external) potential energy V, and the interaction (internal) potential energy U.

The kinetic energy functional

$$T[n_{GS}(r)]$$

can be taken from the Thomas-Fermi approximation or one of its later refinements, and the external potential is simply expressed as

$$V[n_{GS}(r)] = \int d^3r \, V(r)\, n_{GS}(r).$$

The key challenge is to find an accurate interaction energy functional

$U[n_{GS}(r)]$.

Within the framework of Kohn-Sham density functional theory (named after Walter Kohn and Lu Jeu Sham) [24], the intractable many-body problem of interacting electrons in a static external potential is converted to a tractable problem of non-interacting electrons moving in an effective ("Kohn-Sham") potential. The effective potential includes the external potential and the effects of the Coulomb interactions between the electrons, e.g., the exchange and correlation interactions. Modeling the latter two contributions is the chief difficulty. The simplest approximation is the so-called local-density approximation, using the exact Thomas-Fermi exchange energy of a uniform electron gas.

The single-particle Schrödinger equation

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + v_{eff}(r)\right)\phi_i(r) = \varepsilon_i\,\phi_i(r)$$

which involves the fictitious effective local "Kohn-Sham" potential $v_{eff}(r)$ is called the Kohn-Sham equation. As it describes non-interacting fermions, the resulting many-body Kohn-Sham wave function is constructed as a Dirac-Slater determinant from the set of lowest-energy Kohn-Sham orbitals ϕ_i. Thus, the total N-electron density is given by

$$n(r) = \sum_{i=1}^{N} |\phi_i(r)|^2$$

(from now on we shall skip the obvious subscript "GS"). In the following expression for the total energy functional

$$E[n] = T_s[n] + V_{ext}[n] + V_H[n] + E_{xc}[n]$$

The first term T_s is the Kohn-Sham kinetic energy functional, expressed through the Kohn-Sham orbitals as

$$T_s[n] = \sum_{i=1}^{N} \int d^3r\, \phi_i^*(r)\left(-\frac{\hbar^2}{2m}\nabla^2\right)\phi_i(r)$$

.

The second term V_{ext} is the external single-particle potential (e.g., the electron-ion interaction)

$$V_{ext}[n] = \int d^3r\, v_{ext}(r)\, n(r)$$

.

The third term V_H is the Hartree energy

$$V_H[n] = \frac{e^2}{2}\int d^3r \int d^3r' \frac{n(r)\,n(r')}{|r-r'|}$$

.

Finally, the fourth and last term E_{xc} is the exchange-correlation energy.

The Kohn-Sham equations are obtained self-consistently by the variation of the total energy functional E[n] with respect to the set of Kohn-Sham orbitals which in turn define the Kohn-Sham potential in the following way

$$v_{eff}(r) = v_{ext}(r) + v_H(r) + v_{xc}(r),$$

with the Hartree potential

$$v_H(r) = e^2 \int d^3r' \frac{n(r')}{|r - r'|}$$

and the exchange-correlation potential defined as

$$v_{xc}(r) = \frac{\delta E_{xc}[n]}{\delta n(r)},$$

which must be determined independently.

Let us also mention that the Kohn-Sham orbital energies ε_i have little physical meaning in themselves; their connection to the total energy being

$$E = \sum_{i=1}^{N} \varepsilon_i - V_H[n] + E_{xc}[n] - \int d^3r \frac{\delta E_{xc}[n]}{\delta n(r)} n(r).$$

The exchange-correlation energy can be estimated from a local density approximation (LDA). There are a number of choices how to define a local approximation (i.e., one which depends only on the value of the electronic density at a given position, and not on its derivatives) to the exchange-correlation energy. However, most common and successful are those derived from the homogeneous electron gas model (to the extent that LDA is often regarded synonymous with the functionals based on the homogeneous electron gas approximation). The "jellium" model of a homogeneous electron gas involves a large number N interacting electrons in a large space volume V (yielding a finite density n=N/V), with a uniform positive charge compensating background.

A local density approximation for the exchange-correlation energy is generally written as

$$E_{xc}^{LDA}[n] = \int d^3r \; \varepsilon_{xc}(n(r))\, n(r)$$

where $\varepsilon_{xc}(n)$ is the exchange-correlation energy density, a function of electron density alone.

The exchange and correlation contributions are separated

$$E_{xc}^{LDA} = E_x^{LDA} + E_c^{LDA}$$

and sought independently.

In the first term, the exchange energy density ε_x can be calculated analytically for the jellium model of a homogeneous electron gas (compare an equivalent earlier expression for exchange energy per particle, using parameters α and r_s)

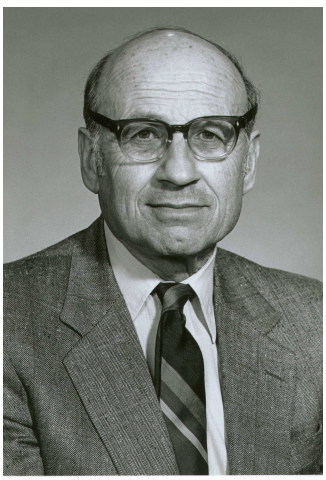
$$\varepsilon_x^{HEG} = -\frac{3e^2}{4}\left(\frac{3}{\pi}\right)^{1/3} n^{4/3},$$

and then (as a local density approximation) used in an unchanged form also for an arbitrary non-homogeneous system, by taking

$$\varepsilon_x(r) = \varepsilon_x^{HEG}(n(r)) = -\frac{3e^2}{4}\left(\frac{3}{\pi}\right)^{1/3} n(r)^{4/3},$$

that is, by assuming that the relation between ε_x and n is local (i.e., independent of $\partial n/\partial r$).

An accurate estimate to the correlation energy density $\varepsilon_c(n)$ poses a bigger problem. Analytic expressions for the correlation energy density of the homogeneous electron gas are not known, except for the high- and low-density limits (corresponding to an infinitely-weak and infinitely-strong correlation, respectively). For intermediate densities, the best known answer is obtained from the quantum Monte Carlo computations [25].



Walter Kohn (1923-)

8. Quantum Monte Carlo

Another known strategy to solve the many-body problem which involves correlations is through a direct search for an explicitly correlated wave-function to minimize the (interacting) many-body Hamiltonian. The arising Hamiltonian matrix elements between many-body states are in form of multi-dimensional integrals, which are computed using a (stochastic) Monte Carlo method. Several variants of the quantum Monte Carlo method have been introduced, known as: variational, diffusion, or path integral quantum Monte Carlo, etc.

The motivation for the quantum Monte Carlo method as opposed to the density functional theory is in that the many-body wave-function ψ, although a complicated mathematical object, obeys a well-known (Schrödinger equation)

$$H\psi = E\psi$$

while the equation which governs the dynamics of the (admittedly, relatively much simpler) total density function n is generally unknown

$$E[n] = T_s[n] + V_{ext}[n] + V_H[n] + E_{xc}[n]; \quad E_{xc}[n] = ?.$$

Thus, arguably, the available exchange-correlation functionals $E_{xc}[n]$ are serious uncontrolled approximations, potentially spoiling the accuracy of the density functional theory approach.

In contrast, quantum Monte Carlo method does not depend on those approximations. However, it is relatively more demanding in terms of the computation costs, and thus only capable of addressing problems with a relatively smaller number of particles. Thus the key issue when comparing quantum Monte Carlo to density functional theory (and, on the other side, to the exact numerical diagonalization) is the scaling of the computation costs of each method with the system size and the accuracy of the representation of a particular physical problem (possibly with an essentially infinite number of particles) by a numerically tractable finite system of N particles. In the last aspect, the boundary conditions may become important as construction of a finite, N-particle model of an extended system characterized by a finite density necessarily involves some kind of a boundary (either edge or periodicity) which might or might not introduce a significant artifact to the model.

Let us begin by considering only real wave functions

$$\Psi(X)$$

expressed through a composite many-body position (r) and spin (σ)

$$X = (\{r_1, \sigma_1\}, \{r_2, \sigma_2\}, \ldots, \{r_N, \sigma_N\}).$$

The numbers of spin-↑ and spin-↓ electrons will be denoted by

$$N_\uparrow \text{ and } N_\downarrow = N - N_\uparrow.$$

As the electrons are fermions, the many-electron wave-function Ψ must be anti-symmetric with respect to an exchange of any pair of electrons. Consider an expectation value of a spin-independent operator A with respect to Ψ

$$\langle A \rangle_{\Psi^2} = \frac{\langle \Psi | A | \Psi \rangle}{\langle \Psi | \Psi \rangle} = \frac{\sum_{\{\sigma\}} \int dR \, \Psi(X) A(R) \Psi(X)}{\sum_{\{\sigma\}} \int dR \, \Psi^2(X)}$$

where

$$R = (r_1, r_2, \ldots, r_N)$$

is a position in the 3N-dimensional "configuration space", and the sums run over all spin configurations with a given N↑ and N↓.

Using antisymmetry of wave function, X can be replaced by

$$X' = (\{r_1, \uparrow\}, \ldots, \{r_{N_\uparrow}, \uparrow\}, \{r_{N_\uparrow+1}, \downarrow\} \ldots, \{r_N, \downarrow\}),$$

which allows relabeling components of the dummy integration variable and taking advantage of the fact that integrals for each spin configuration {σ} with the same numbers N↑ and N↓ are the same. Thus, the spin sums can be removed, and the expectation value can be evaluated as

$$\langle A \rangle_{\Psi^2} = \frac{\int dR \, \Psi(R) A(R) \Psi(R)}{\int dR \, \Psi^2(R)}$$

where

$$\Psi(R) = \left(\{r_1, \uparrow\}, \ldots, \{r_{N_\uparrow}, \uparrow\}, \{r_{N_\uparrow+1}, \downarrow\} \ldots, \{r_N, \downarrow\}\right)$$

is (only) anti-symmetric with respect to the exchanges of electrons having the same spin (as the electrons with different spin are effectively distinguishable).

The normal starting point for the construction of a quantum Monte Carlo many-electron trial wave function is the set of single-electron orbitals from either Hartree-Fock approximation or density functional theory. An anti-symmetric wave-function is constructed out of such orbitals in a standard manner, by forming a Dirac-Slater determinant. When electrons are unpolarized, a product of anti-symmetric wave-functions (a pair of determinants, separately for spin-↑ and spin-↓ electrons) is constructed instead of a single determinant

$$\Psi(R) = D_\uparrow(r_1, r_2 \ldots, r_{N_\uparrow}) D_\uparrow(r_{N_\uparrow+1}, r_{N_\uparrow+2} \ldots, r_N)$$
$$= \begin{vmatrix} \psi_1^\uparrow(r_1) & \cdots & \psi_{N_\uparrow}^\uparrow(r_1) \\ \vdots & & \vdots \\ \psi_1^\uparrow(r_{N_\uparrow}) & \cdots & \psi_{N_\uparrow}^\uparrow(r_{N_\uparrow}) \end{vmatrix} \begin{vmatrix} \psi_1^\downarrow(r_{N_\uparrow+1}) & \cdots & \psi_{N_\downarrow}^\downarrow(r_{N_\uparrow+1}) \\ \vdots & & \vdots \\ \psi_1^\downarrow(r_N) & \cdots & \psi_{N_\downarrow}^\downarrow(r_N) \end{vmatrix}.$$

Here, ψ are the set of lowest-energy single-electron orbitals. Of course, using two smaller determinants instead of one is advantageous because computation of a determinant costs ~N^3.

The main advantage of quantum Monte Carlo is the capability of using more complicated wave functions than Dirac-Slater determinants. Furthermore, the correlation effects are introduced in compact fashion through the adequate prefactors (in contrast to the post-Hartree-Fock methods using computationally inconvenient sums of many determinants).

Thus, the starting-point wave-function in form of a Dirac-Slater determinant is improved by adding a so-called Jastrow prefactor J(R) expressed in terms of relative coordinates and thus explicitly accounting for the correlations among the individual electrons

$$\Psi(R) = \exp[J(R)]\, D_\uparrow(r_1, r_2 \ldots, r_{N_\uparrow}) D_\uparrow(r_{N_\uparrow+1}, r_{N_\uparrow+2} \ldots, r_N).$$

Since the Dirac-Slater determinant is anti-symmetric with respect to pair electron exchanges, the Jastrow prefactor must be symmetric (to maintain the correct anti-symmetry of the whole wave-function). Thanks to the exponential form of the Jastrow prefactor in the wave function, J(R) can be written as a sum of different terms corresponding to a number of different causes of correlation (e.g., among the electrons, among an electron and a nucleus, among a pair of electrons and a nucleus, etc.; terms describing correlations of electrons with the same or opposite spins are also conveniently distinguished).

The Jastrow prefactor must satisfy certain general boundary conditions called the Kato cusp conditions. These are (for r_{ij} , r_i and r_j all treated as independent variables)

$$\left(\frac{\partial J}{\partial r_{ij}}\right)_{\substack{r_{ij}=0 \\ r_i=r_j}} = \frac{1}{2}$$

for the electron pair of opposite spin (↑↓), and

$$\left(\frac{\partial J}{\partial r_{ij}}\right)_{\substack{r_{ij}=0 \\ r_i=r_j}} = \frac{1}{4}$$ for the electron pair of the same spin (↑↑).

Note that the addition of the Jastrow factor does not affect the nodal surface (the set of points R in the configuration space for which the wave function vanishes) which is completely determined by the constituent Dirac-Slater determinant.

In variational quantum Monte Carlo (VMC), the many-body wave function contains a number of free parameters which are optimized using the variational principle (e.g., by minimizing expectation value of Hamiltonian). The VMC energy is therefore defined as the expectation value of the Hamiltonian H with respect to trial wave function Ψ. It can be written as

$$\langle H \rangle = \frac{\int dR\ E_L(R)\,\Psi^2(R)}{\int dR\ \Psi^2(R)}$$

where

$$E_L(R) = \Psi^{-1}(R) H\ \Psi(R)$$

is called the "local energy." The VMC energy can be conveniently evaluated by using the Metropolis algorithm to generate a sequence of configurations {R} distributed according to Ψ^2(R) and then by averaging the corresponding local energies. If needed, the expectation values of other operators are evaluated in a similar fashion.

9. Configuration interaction metod

The method of exact diagonalization of the Hamiltonian matrix in the basis of single-particle configurations (called the configuration interaction method) will be described in detail in the following sections. In brief, it consists of spanning the many-body Hilbert space by Dirac-Slater determinants (or "configurations") which involve a convenient set of single-particle spin-orbitals (for example, although not necessarily, the eigenstates of the single-particle Hamiltonian h). In this basis, the Hamiltonian matrix often becomes sparse, since the pair interaction (typically representing the off-diagonal part of the Hamiltonian matrix) can only couple (through an elementary process of the two-electron scattering) those configurations which differ by the occupation of at most four spin-orbitals. Such matrix can be efficiently diagonalized using various iterative methods related to the classical Lanczos algorithm. If only the actual sought many-body state indeed falls (to sufficient projection) in the chosen configuration basis, the method is essentially exact in describing the correlation effects. Unfortunately, high computation cost limits its application to smaller systems than those tractable by quantum Monte Carlo (or density functional theory).

III. Configuration interaction method

1. Example of exact diagonalization (single particle discretization)

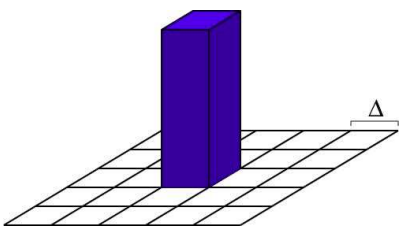
1.1 Schrödinger equation in discretized real space

As an example of application of the method of exact diagonalization (although not yet in the configuration basis) let us consider a single particle moving in an external potential V(r). As a basis of states, let us use the states localized in the narrow areas of the real space. To do this, let us discretize the space in each direction (x, y, …) by introducing a rectangular grid. For simplicity, the total length of the system in each direction will be assumed equal and called L. The number of divisions in each direction will also be assumed equal, such that each space box has the dimensions Δ^d, where d is the dimensionality of the allowed space. The dimension of basis is therefore $N=(L/\Delta)^d$, and we shall assume that N can be sufficiently large that the resulting N×N Hamiltonian matrix cannot be diagonalized in a completely straightforward manner.

The basis wave function corresponding to the particle located at box number j is

$\phi_j(r) = \Delta^{-d/2}$ inside the j^{th} box (around $r=r_j$), and

$\phi_j(r) = 0$ elsewhere.



We shall also use the standard “bra-ket” notation

$\phi_j = |j\rangle$

Obviously, the boxes do not overlap, so the basis is orthonormal

$\langle i|j\rangle = \delta_{ij}$.

Strictly speaking, being discontinuous at the box edge, these are not valid wave functions. However, they still allow to obtain a computational scheme which gives the correct physical results in the limit of $\Delta \to 0$.

The Hamiltonian operator contains the kinetic energy K and the external potential V. Its matrix elements in the above basis are (we shall use the effective units, with the Planck constant and the mass both set to unity)

$$K_{ij} \equiv \langle i|K|j\rangle = \frac{1}{2}\int d^d r\, \phi_i^*(r)\,\nabla^2\phi_j(r)$$

and

$$V_{ij} \equiv \langle i|V|j\rangle = \int d^d r\, \phi_i^*(r)\,V(r)\,\phi_j(r) = \delta_{ij}\int d^d r\, V(r)\,|\phi_i(r)|^2 \approx \delta_{ij}V(r_i)\,.$$

For the kinetic energy, derivatives can be approximated by the central difference operator. In one dimension (easily generalized to d>1) this is

$$\begin{aligned}\Delta^{-2}\,\delta^2\phi_j(r) &\equiv \Delta^{-2}\left[\,\phi_j(r-\Delta) - 2\phi_j(r) + \phi_j(r+\Delta)\,\right] \\ &= \Delta^{-2}\left[\,\phi_{j-1}(r) - 2\phi_j(r) + \phi_{j+1}(r)\,\right]\end{aligned}$$

so that the kinetic energy matrix element becomes

$$K_{ij} = \frac{1}{2}\int d^d r\, \phi_i^*(r)\,\Delta^{-2}\delta^2\phi_j(r) = \begin{cases} -\Delta^{-2}/2 & \text{for } j = i\pm 1 \\ \Delta^{-2} & \text{for } j = i \end{cases}\,.$$

Or, the action of the kinetic energy operator on a basis state can be written as

$$K|i\rangle = -\Delta^{-2}\left[\frac{1}{2}|i-1\rangle - |i\rangle + \frac{1}{2}|i+1\rangle\right].$$

Combining both kinetic energy and external potential contributions, the only non-vanishing matrix elements of the Hamiltonian are

$$H_{i,i} = V(r_i) + \frac{1}{\Delta^2}\,,\ \text{and}$$

$$H_{i,i\pm1} = -\frac{1}{2\Delta^2}\,.$$

Generalization to more than one dimension (d>1) is straightforward

$$H_{i,i} = V(r_i) + \frac{d}{\Delta^2}\,,\ \text{and}$$

$$H_{i,\delta[i]} = -\frac{1}{2\Delta^2}\,,$$

where δ[i] denotes the 2d nearest neighbours of the box number i. Note that the hopping matrix element becomes stronger for the smaller volume elements (in accordance with the uncertainty principle).

1.2 Continuum limit

Let us now demonstrate the correct continuum limit of this model, on an example of a free particle in one dimension, in a periodic box (so-called closed chain). The correct energy eigenstates are

$$\psi_k(r) = e^{ikr}$$

with k=2πn/L, n=0, 1, …; the corresponding eigenenergies are

$$E_k = \frac{1}{2}k^2$$

.

Let us now consider this problem in the discretized space. The discrete coordinate

$$r_j = j\Delta = jL/N$$

puts a limit on the allowed (inequivalent) momenta. Namely, since

$$\exp\left[i\left(k+\frac{2\pi}{\Delta}\right)r_j\right] = \exp\left[ikr_j\right]\exp\left[i\frac{2\pi}{\Delta}j\Delta\right] = \exp\left[ikr_j\right]$$

There only allowed momentum values are

$$k_n = n\frac{2\pi}{L} = \frac{n}{N}\frac{2\pi}{\Delta}$$

.

Let consider the following proposed state

$$|k\rangle = \frac{1}{\sqrt{N}}\sum_{j=0}^{N-1} e^{ikr_j}|j\rangle$$

and determine the action of the Hamiltonian (i.e., the kinetic energy operator) on this state

$$H|k\rangle = \frac{1}{\sqrt{N}}\sum_{j=0}^{N-1} e^{ikr_j}H|j\rangle = -\frac{1}{\Delta^2}\frac{1}{\sqrt{N}}\sum_{j=0}^{N-1} e^{ikr_j}\left[\frac{1}{2}|j-1\rangle - |j\rangle + \frac{1}{2}|j+1\rangle\right]$$

.

After shifting the dummy indices in the j±1 terms, this becomes

$$H|k\rangle = -\frac{1}{\Delta^2}\frac{1}{\sqrt{N}}\sum_{j=0}^{N-1} e^{ikr_j}\left[\frac{1}{2}\left(e^{ik\Delta}+e^{-ik\Delta}\right)-1\right]|j\rangle = -\frac{1}{\Delta^2}\left(\cos k\Delta - 1\right)|k\rangle$$

.

Hence, |k⟩ has been demonstrated to be an eigenvector, with a corresponding eigenenergy

$$E_k = \frac{1}{\Delta^2}\left(1-\cos k\Delta\right)$$

which for a small Δ can be expanded into the following Taylor series

$$E_k = \frac{1}{2}k^2 - \frac{1}{24}\Delta^2 k^4 + \ldots$$

which agrees to the leading order with the continuum result, and therefore justifies the discretization procedure as a treatment of the kinetic energy. Note here also that the discretized kinetic energy is lower than the true energy.

1.3 Diagonalization

The above established procedure allows conversion of a problem of a particle moving in a continuous d-dimensional space in the field of an external potential – into the problem of a sparse N-dimensional Hamiltonian matrix, which can be diagonalized quasi-exactly (on a computer).

Since the matrix is large and sparse, and since most likely only some of the low-energy eigenenergies and/or eigenstates are needed as physically meaningful, a particularly suitable diagonalization method is the Lanczos algorithm (which will be discussed in detail later). The key idea (common to all "power" methods) is that a repeated action of a Hamiltonian on an arbitrary state Φ (which can be expanded into a linear combination of the eigenstates Ψ_k) creates a state whose expansion is dominated by those eigenstates which correspond to the extremal eigenvalues).

2. Lanczos algorithm

Let us consider a large Hamiltonian matrix H. The following procedure converts it into the tridiagonal form, and thus transforms the original arbitrary dynamical problem into a one-dimensional chain in which each subsequent state is only coupled to its pair of backward and forward neighbors.

1. Let us begin with a normalized vector $\mathbf{u}_0$ (either guessed or chosen at random).
2. Vector $\mathbf{v}_1$ is obtained by action of the Hamiltonian onto $\mathbf{u}_0$
   $v_1 = Hu_0$.
3. Then, vector $\mathbf{w}_1$ is obtained by orthogonalization of $\mathbf{v}_1$ to $\mathbf{u}_0$
   $w_1 = v_1 - (v_1 \bullet u_0)\, u_0$.
4. Finally, the next Lanczos vector $\mathbf{u}_1$ is obtained by the normalization of $\mathbf{w}_1$
   $u_1 = \dfrac{w_1}{|w_1|}$.

The preceding steps are now repeated so that the next Lanczos vector $\mathbf{u}_{n+1}$ is constructed by acting with H onto the preceding Lanczos vector $\mathbf{u}_n$, followed by orthogonalization to *two* preceding Lanczos vectors and finally completed by the normalization

5. $v_{n+1} = Hu_n$;
6. $w_{n+1} = v_{n+1} - (v_{n+1} \bullet u_{n-1})\, u_{n-1} - (v_{n+1} \bullet u_n)\, u_n$;
7. $u_{n+1} = \dfrac{w_{n+1}}{|w_{n+1}|}$.

Thus, beginning with a (nearly) arbitrary starting vector $\mathbf{u}_0$, the above procedure generates the so called Lanczos basis

$\{u_n\}$

with the following properties:

1. Hamiltonian is tridiagonal in the Lanczos basis
   (despite orthogonalization only to two preceding vectors in the procedure).
2. The Lanczos matrix elements
   $H_{n,n} = \mathbf{v}_{n+1} \bullet \mathbf{u}_n$
   $H_{n,n-1} = \mathbf{v}_{n+1} \bullet \mathbf{u}_{n-1}$
   are computed during the basis generation (with no additional computational cost).
3. The dominant cost of each Lanczos iteration is one matrix-vector multiplication
4. When analyzed as a function of Lanczos dimension (i.e., the number of generated Lanczos vectors; i.e., the number of Lanczos iterations), the extreme Lanczos eigenvalues (eigenvalues of the Hamiltonian expressed in the Lanczos basis) converge exponentially quickly to the exact eigenvalues of the Hamiltonian (in the full basis).
5. The number of Lanczos iterations needed to reach full numerical convergence for the ground state is typically much smaller than the dimension of the Hamiltonian matrix.
6. The cost of diagonalization of the tridiagonal Lanczos matrix is of the order of N operations (i.e., negligible compared to the cost of matrix-vector multiplication performed during each Lanczos iteration).

The following should be realized when chosing the initial Lanczos vector $\mathbf{u}_0$:

1. Lanczos procedure preserves the symmetry of the initial vector $\mathbf{u}_0$
   (at least, in principle, assuming lack of accumulating numerical errors).
2. In practice, accumulation of the numerical errors (after a sufficient number of iterations) may lead to an escape from the invariant subspace of $\mathbf{u}_0$ (of course, if only such escape should lower the energy).
3. Still, if the symmetry of the sought ground state is known, choosing $\mathbf{u}_0$ with this symmetry is desirable so as to reduce the number of iterations until convergence.
4. Also, choosing $\mathbf{u}_0$ with a large overlap with the eventual ground state improves the convergence.
5. Lacking such insight, $\mathbf{u}_0$ can be generated at random and still guarantee convergence with the probability of one.

Observations regarding the computational cost:

1. The dominant cost is a single matrix-vector multiplication in each iteration.
2. This multiplication can be optimized for efficiency by taking advantage of the structure of the matrix (sparsity).
3. The number of iterations needed to reach convergence if usually of the order of 10^2.
4. Since the matrix-vector multiplication can be done “on-the-fly” and since only three Lanczos vectors need be stored in memory at each given time, the dimensions tractable on the currently available PC workstations exceed one billion. For example, for the dimension of $D=10^9$, the memory required to store three vectors is 24GB.

Sparsity of the Hamiltonian is essential for an efficient matrix-vector multiplication. It occurs naturally for a pair interaction expressed in the configuration interaction basis. In order to demonstrate it, consider a Hamiltonian with a two-body interaction V

$$\hat{H} = \sum_i T_i + \sum_{i,j} V_{ij}$$

which in the second quantization reads

$$\hat{H} = \sum_{i,j} T_{ij}\ a_i^+ a_j + \sum_{i,j,k,l} V_{ijkl}\ a_i^+ a_j^+ a_k a_l$$

(where a and a+ are the annihilation and creation operators).

Consider also the configuration basis for N particles

$$|i\rangle = |i_1, i_2, \ldots, i_N\rangle = a_{i_1}^+ a_{i_2}^+ \ldots a_{i_N}^+ |vac\rangle .$$

It is essential to realize that the matrix element of the two-body interaction Hamiltonian H in the basis of single-particle configurations

$$\langle i|\hat{H}|j\rangle$$

must vanish unless these two configurations |i⟩ and |j⟩ differ only at four or fewer orbitals, so that they can be connected with a quartic term like $(a^+a)(a^+a)$. This fact naturally produces a sparse matrix, increasingly so for the larger N (in comparison to 2).

3. Manipulation and storage of sparse matrices

Since the Hamiltonian matrices suitable for exact diagonalization are commonly large and sparse, there are essentially two strategies to deal with the problem of their storage: One is the compression; the other is avoiding storage by carrying out the matrix-vector multiplication "on-the-fly". Let us begin with the concept of matrix compression.

The basic idea is to avoid storing zero matrix elements (and possibly also repeated elements). In the most direct method, the position (row and column) as well as the value are stored for each of the nonzero matrix elements. Let us denote the matrix dimension by D. Then the memory required to store the whole uncompressed matrix in the standard "double" precision (8 bytes per floating point number) is

$$8B \cdot D^2 .$$

When compressed, this becomes

$$(2 \cdot 4B + 8B) \cdot G = 16B \cdot G$$

where it was assumed that a four-byte integer is sufficient/needed to store the row or column index. Defining the sparsity s as the fraction

$$s = G/D^2$$

the compressed-to-uncompressed memory ratio becomes

$$\frac{16B \cdot G}{8B \cdot D^2} = 2s$$

And it hugely favours compression for the sparse matrices (s<<1).

Further obvious compression can take advantage of the fact that the matrix may be symmetric (and this only a half, either one above or below the diagonal, need be stored) or that the diagonal may have a very different sparsity then the off-diagonal part (and thus the two can be stored separately).

Additional compression can employ the fact that many matrix elements may repeat themselves, and thus they may be tabulated so that each non-zero element would be stored with a short (possibly two-byte) pointer to the table of all genuine matrix elements rather than directly to a specific eight-byte value. Such repetition of matrix elements naturally occurs for the pair interaction Hamiltonians in the configuration basis, where the scattering matrix element between two N-body configurations depends only on those orbitals defining the two-body scattering process that couples the N-body configurations, and the occupation of any other orbitals (unchanged in this two-body scattering can only affect the sign of the matrix elements (through the anti-commutation rules for fermions)

$$\langle i_1, i_2, \ldots, A', \ldots, B', \ldots, i_N | H | i_1, i_2, \ldots, A, \ldots, B, \ldots, i_N \rangle = \pm \langle A', B' | H | A, B \rangle .$$

Thus, the magnitude of the above matrix element remains insensitive to the (unchanged) N-2 occupied orbitals, and this element is bound to repeat in the matrix for their any other combination – as long as (A,B,A',B') remain the same.

A particularly convenient compression format (for the purpose of efficient matrix-vector multiplication) is the so-called "compressed-row format." In this scheme, the non-zero matrix elements are stored row by row, so that for each of them only the value (or a pointer to the table of values) and the column need be stored. Additionally, it must be stored which matrix element begins a new row so that the complete position (row and column) can be restored for each matrix element value.

As an example, let us consider the following symmetric matrix

$$H = \begin{bmatrix} 4.5 & 0 & 1.1 & 5.2 & 0 \\ 0 & 2.0 & 0 & 3.2 & 3.3 \\ 1.1 & 0 & 1.5 & 0 & 0 \\ 5.2 & 3.2 & 0 & 7.2 & 6.6 \\ 0 & 3.3 & 0 & 6.6 & 4.0 \end{bmatrix}$$

Let us call the vector of diagonal matrix element values as DH[...],the vector of off-diagonal matrix element values as OH[...], the vector of columns as JH[...], and the vector of new-row-markers as IH[...].

$$DH = [4.5, 2.0, 1.5, 7.2, 4.0]$$
$$OH = [1.1, 5.2, 3.2, 3.3, 6.6]$$
$$JH = [3, 4, 4, 5, 5]$$
$$IH = [1, 3, 5, 5, 6]$$

The length of both DH and IH is the dimension of the matrix D, and the length of both OH and JH is the number of non-zero matrix elements above the diagonal G. Using this compression, the vector-matrix multiplication

$$Q = HP \quad \Leftrightarrow \quad \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ Q_5 \end{bmatrix} = \begin{bmatrix} 4.5 & 0 & 1.1 & 5.2 & 0 \\ 0 & 2.0 & 0 & 3.2 & 3.3 \\ 1.1 & 0 & 1.5 & 0 & 0 \\ 5.2 & 3.2 & 0 & 7.2 & 6.6 \\ 0 & 3.3 & 0 & 6.6 & 4.0 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{bmatrix}$$

proceeds as follows (in Fortran)

```
DO row = 1, D
     Q[row] = P[row] * DH[row]
END DO

DO row = 1, D-1
     DO element = IH(row), IH(row+1)-1
          col = JH[element]
          val = OH[element]
          Q[row] = Q[row] + P[col] * val
          Q[col] = Q[col] + P[row] * val
     END DO
END DO
```

Furthermore, if the matrix elements within each row are ordered by column, then instead of storing columns in vector JH, one may choose to store the increase in the column index compared to the previous matrix element in the same row. As these increments will generally be smaller than the columns themselves, storing them in JH may require fewer bytes per entry (e.g., two instead of four). As a safety net against (rare) situations when the increment is so large that it will not fit in the allocated memory cell (e.g., two bytes), one may add another vector JJH of 4-byte integers (and placing a conventional marker such as "-1" into JH would signal that a given column need be read from JJH.

4. Efficient computation of matrix elements

Speed- and memory-efficient computation of the Hamiltonian matrix is especially crucial when it is calculated "on the fly" instead of being stored between successive Lanczos iterations. This is achieved by a rather literal implementation of the second quantization, as will be described below.

4.1. Representation of the configuration basis

Each basis state (N-particle configuration)

$$|i\rangle = |i_1, i_2, \ldots, i_N\rangle = a_{i_1}^{+} a_{i_2}^{+} \ldots a_{i_N}^{+} |vac\rangle$$

is *not* stored as (what might appear as a natural choice) a table of indices

$index\left[k=1\ldots N, i=1\ldots D\right]$

where a composite "index" denotes the single-particle state created by a^{+}_{index}, k labels the particles, and i labels the configurations. Such storage depends explicitly on N (which requires explicit scaling of the programs), by explicit labelling the particles it does not take advantage of the anti-symmetrization of the configurations (in fact, to avoid their double counting, the genuine configurations have to be generated with an additional requirement that (for example) the indices are ordered as follows

$index\left[k,i\right] \prec index\left[k',i\right] \quad \text{for} \quad k<k'$

which is also somewhat problematic for composite indices. Furthermore, storage in this way is not memory-efficient, by taking

$4B \cdot N \cdot D$

of space for the whole basis (assuming that each index can be stored as a 4-byte integer).

Therefore, a better option is to store each basis state (configuration) as a sufficiently long binary number

$$00101100101100010 1\ldots$$

where the position of each bit is associated with a single-particle spin-orbital, and the value of the bit (1 or 0) represents that a given spin-orbital is occupied or not.

For example, the four-particle configuration

$\left|0,3,4,6\right\rangle = a^{+}_{0}\, a^{+}_{3}\, a^{+}_{4}\, a^{+}_{6} \left|vac\right\rangle$

would be represented by a binary number

$10011010000\ldots = 2^{0}+2^{3}+2^{4}+2^{6}=89$

For this coding, the length of the binary number need be sufficient to store all spin-orbitals. For example, a convenient standard four- and eight-byte integers allow for 32 or 64 spin-orbitals, respectively; a larger number of spin-orbitals in a given problem may require a somewhat less convenient coding in form of several integers. In contrast to the previous coding, the total memory needed to store a basis scales with the number of spin-orbitals S rather than with the number of electrons

$1B \cdot S/8 \cdot D$

In practice, when using standard (4-byte) or long (8-byte) integers instead of S-bit long ones, this memory amounts to

$4B \cdot D \quad \text{or} \quad 8B \cdot D.$

However, the biggest advantage is not in the potential savings in memory space, but in the elegance and ease to compute the many-body matrix elements.

For the reasons that will become clear in the next paragraphs, it is convenient to sort the basis so that the binary numbers representing the consecutive configurations are ordered.

4.2. Finding non-zero matrix elements

In the computation of the Hamiltonian matrix (or the matrix of any other relevant operator) it is critical to efficiently determine those elements which are non-zero. To this end, the arguably natural method of scanning the matrix element by element and checking which one is potentially non-zero (for example, by checking which Hamiltonian matrix element connects two configurations which are only different in at most two spin-orbitals, as allowed by a two-body interaction scattering) – is dramatically inefficient. Indeed, it requires

$$D^2$$

such checks, which can be dramatically reduced by scanning only the rows of the matrix (that is, going through the matrix row by row) or, equivalently, going through the entire basis configuration by configuration. For each such "initial" configuration, all possible "final" configurations are generated by literally applying the second-quantization version of the Hamiltonian. For the Hamiltonian including a two-body interaction this means

$$\hat{H}\left|i_1,i_2,\ldots,i_N\right\rangle = \sum_{i,j} T_{ij}\ a_i^+ a_j \left|i_1,i_2,\ldots,i_N\right\rangle + \sum_{i,j,k,l} V_{ijkl}\ a_i^+ a_j^+ a_k a_l \left|i_1,i_2,\ldots,i_N\right\rangle$$

where each non-zero element in the sums is a different allowed "final" configuration, with up to two spin-orbitals which are different from the "initial" configuration, as given by the one-body scattering j←i or the two-body scattering (i,j)←(k,l). The absolute magnitudes of matrix elements constructed in this way are obviously T_{ij} or V_{ijkl} (explicitly only dependent on those spin-orbitals between which the scattering actually occurs), respectively, so they do not require additional computation. However, the signs in front of these matrix elements does depend on the occupation of all other spin-orbitals, as follows from the anti-commutation of the fermionic creator operators a^+ in the definition of the configuration

$$\left|i\right\rangle = \left|i_1,i_2,\ldots,i_N\right\rangle = a_{i_1}^+ a_{i_2}^+ \ldots a_{i_N}^+ \left|vac\right\rangle .$$

A simple rule is that once a certain (arbitrary) ordering of spin-orbitals is assumed and used consistently, each scattering process in which all (one or two) scattering particles jump over a total number of occupied orbitals which is even – has the amplitude (matrix element) T_{ij} or V_{ijkl}, while the scattering processes in which all (one or two) scattering particles jump over a total number of occupied orbitals which is odd – have the amplitudes (matrix elements) with a reversed sign: $-T_{ij}$ or $-V_{ijkl}$.

The number of operations needed to compute the whole matrix is now proportional to

$$D^1$$

and also to the number of allowed one- and two-body scatterings (which is of the same order of magnitude as the number of operations in a single "check" of the previous method) and the number of operations needed to determine the column of each matrix element.

Indeed, let us carefully observe that while in the method of scanning the entire matrix element by element both row and column of each element are determined automatically, in the present method of running through the initial configurations it costs additional effort to determine the column. Since the column of the matrix element is simply the position of the corresponding "final" configuration in the basis, determination of the column is equivalent to the search of a given "final" configuration in the basis.

Now, in the most primitive search procedure which involves comparison of the given "final" configuration with all basis configurations one by one, this gives an additional factor of D^1 operations per matrix element so that altogether the method also involves

$$D^2$$

of "check" operations.

However, if the basis is sorted (and the binary representation of the basis configurations allows their convenient and efficient comparison with respect to position in the basis), then the search can be done much faster, for example by the method of bisection (note that bisection does not require that the position-to-configuration function is continuous, only that it is monotonic) which scales as log D, so that the total number of operations scales as

$$D \log D$$

which, obviously, can be orders of magnitude lower than D^2.

4.3. Determining the column of a matrix element

Let us look more closely at the application of bisection method to the calculation of the column of a matrix element.

Let us denote the position of each configuration in the basis by

$$j = 1, 2, \ldots, D$$

and their (binary) value of each configuration by

b.

The function

$$b(j)$$

is precisely the definition about the basis which is stored in memory (this is simply the information about the occupied spin-orbitals in each jth configuration). It can be stored efficiently as a table of elements b, because the arguments j fill densely (actually, even completely) their interval from j=1 to D, and thus they can be treated as indices of the table b. The fact the basis is ordered means that the function $b(j)$ is monotonic.

The calculation of the Hamiltonian matrix begins with the loop over "initial" configurations, that is over j=1…D. For each j, the corresponding configuration b is read immediately from the table $b(j)$ which is stored in memory. The "final" configurations are now also readily constructed from the initial configuration b, which explicitly tells about the occupied and empty spin-orbitals – available for the "scattering from" or "scattering to" processes. However, each "final" configuration b' constructed from the given b by action of the second-quantization Hamiltonian need also be assigned its position j' in the basis, which is equivalent to the column of the given matrix element. This assignment $b' \rightarrow j'$ involves the inverse "basis" function

$$j(b).$$

Unfortunately, the inverse basis function $j(b)$ cannot be stored because the values b are sparse in their relevant interval of all binary numbers with the number of digits dictated by the

number S of allowed spin-orbitals. For example, when the number of particles is N=10 and the number of spin-orbitals is S=30, the total number of configurations (ignoring any symmetries such as connected to the conservation angular momentum, parity, etc.) is

$$D = \binom{S}{N} \approx 3 \cdot 10^7$$

which only takes 4B*D=120MB of memory to store the basis as a table of 4-byte (i.e., 32-bit) integers. In contrast, if the inverse function $j(b)$ were to be stored as a table, its length would have to be the total number of S-digit binary numbers, which is

$$2^S \approx 10^9$$

and would require a much larger (though not yet forbidding) memory of $4B*2^S$=4GB. For a somewhat larger system of the same particle-to-orbital proportion (N,S)=(12,36), storage of b(j) as a table of 8-byte integers is still possible by requiring about 10GB of memory, but storing j(b) as a table of 4-byte integers would require as much as 275GB of memory, not to be found on any widely available workstation.

This comparison of the memory needed to store b(j) and j(b) is even more dramatic when additional symmetries of the Hamiltonian allow its block diagonalization already at the level of the generation of the basis, thus greatly reducing D of each block compared to (S,N).

When j(b) cannot be stored as a table, it must be computed "on the fly". Provided that j(b) is monotonic (that is, the basis is ordered, which can be either ensured during basis generation or achieved by a sorting procedure), this can be done by the method of bisection. Namely, the pair of indices:

$$j_{LO} \quad \text{and} \quad j_{HI}$$

are initially chosen such that

$$b(j_{LO}) < b' < b(j_{HI})$$

and then in each successive iteration the search range is narrowed by replacing either end j_{LO} or j_{HI} by the midpoint

$$j_{MID} = \frac{1}{2}(j_{LO} + j_{HI})$$

depending on whether

$$b' > b(j_{MID}) \quad \text{or} \quad b' < b(j_{MID}).$$

In the absence of a hint, the initial range must be chosen as the full range

$$j_{LO} = 1 \quad \text{and} \quad j_{HI} = D$$

and the number of required iterations is

$$\log_2 D.$$

However, a strategy which is intermediate between storing the entire table j(b) and running bisection in the full range (1,D) can also be easily implemented. In this strategy, the inverse

basis function j(b) is stored partially, in the following manner. The 2^S-long full range of configurations *b* is divided into a number L of sub-intervals, each of length

$$\Lambda = \frac{2^S}{L},$$

and stored in a table are the lowest indices j of all configurations falling into a given sub-interval. Thus, by calculating for a given configuration *b* its corresponding sub-interval index

$$\beta = \frac{b}{\Lambda}$$

and reading from the stored table the lowest indices j_0 of this and of the next sub-interval, the bisection can be restricted to a shorter *j*-interval (limiting the average number of iterations)

$$j_{LO} = j_0(\beta) \quad \text{and} \quad j_{HI} = j_0(\beta + 1) - 1$$

It is particularly convenient and intuitive to choose the sub-interval number L as a power of 2

$$L = 2^K.$$

In this case, the calculation of β is a fast "binary shift" operation (by K bits to the left), and the division into b-sub-intervals simply means the classification of each configuration *b* by its *K* leading bits. In other words, all configurations beginning with a given *K*-bit number have assigned (and stored) a j-sub-interval within which the bisection is carried out. This is clearly an intermediate choice between:

(i) complete classification of configurations (by all of their bits) – that is, storing the entire inverse basis function j(b) in memory and thus eliminating the need for bisection altogether;

(ii) no classification whatsoever – that is running the bisection in the full range (1,D).

The flexibility in choosing K allows convenient optimization of the entire matrix computation procedure with respect to speed and required memory.

5. "Ghost" states

In the absence of numerical errors, Lanczos procedure with explicit orthogonalization to only two preceding Lanczos vectors produces an exactly tri-diagonal Hamiltonian in the Lanczos basis which is exactly orthogonal. However, the accumulation of such (even very small) errors in real computation leads to an abrupt loss of accuracy after a sufficiently large number of operations (that is, iterations).

An example of this effect are the "ghost states", which typically appear in form of a sudden (as a function of the number of Lanczos iterations) drop of the energy of an earlier converged (or nearly converged) excited level. The mechanism is that a sudden loss of orthogonality of the Lanczos basis in effect of an accumulation of small rounding errors results in an artificial doubling of the extremal eigenvalues. Thus, an energy level which suddenly appears in the gap between the actual eigenenergies, after it has departed from a higher, nearly converged level, but before it has converged again to a lower, now artificially doubled level – is called a "ghost state" or "ghost level".

While the “ghost states” have no effect on the convergence of the single lowest (or highest) eigenenergy, they undermine counting of the ground state degeneracy. More importantly, they often make it impossible to reach simultaneous and satisfactory convergence of several excited states, and thus to determine at least the low-energy part of the excitation spectrum.

What are the available remedies for “ghost states”? A fairly obvious one is performing an explicit orthogonalization of each new Lanczos vector to all of the preceding vectors (instead of only to the last two). This option has its costs. First of all, it requires storing all of the preceding Lanczos vectors, either in memory or on the disk. For huge dimensions of the basis, this may easily limit the number of iterations due to the lack of memory. On the other hand, the computational cost of orthogonalization grows steadily with each iteration. To remedy these two effects, the Lanczos procedure can be terminated before the full convergence, and then new initial Lanczos vector or vectors can be constructed from those of the terminated procedure. When not only the ground state but also a number of excited states are sought for, feeding only the ground state of the terminated Lanczos tri-diagonal Hamiltonian as a new initial Lanczos vector – is not sufficient. Instead, a block of an appropriate number of vectors must be obtained and fed in at the start of a new run of the Lanczos algorithm.

6. Calculation of eigenvectors

The need for storage of all Lanczos vectors (instead of only three at a time) arises also when the ground state vectors are sought in addition to the ground state energy. Since, obviously, diagonalization of the tri-diagonal Lanczos Hamiltonian produces a ground state expressed in the Lanczos basis, this basis must be available to subsequently perform the conversion of this ground state to the initial configuration basis. When storing enough vectors is not possible, an alternative is to simply run the Lanczos procedure twice – the first time to build the Lanczos Hamiltonian matrix, whose diagonalization provides the coordinates of the ground or relevant excited states in the Lanczos basis; and the second time – to build the ground state vector “on the fly”, by taking advantage of the earlier determined expansion coefficients.

7. Simultaneous resolution of an additional quantum number

Quite commonly, the goal of diagonalization procedure is not finding the absolute ground state, but the ground state in a particular invariant subspace. For example, the energy spectrum may be sought as a function of an additional conserved quantum number, which defines the splitting of the total Hilbert space into the invariant subspaces corresponding to different eigenvalues of this additional conserved quantity (for example, total parity, angular momentum, or spin). Or, at least, knowledge of the lowest state at a given total spin or angular momentum may be desired in order to compare the two phases (or resolve their competition). Even if only the absolute ground state is sought, running the Lanczos algorithm inside its invariant subspace may speed up convergence as it is likely to enlarge the energy gap above the ground state (the gap to the next state of the same symmetry likely being larger than the gap to the next state of any symmetry whatsoever), and the convergence rate of the Lanczos procedure is known to depend critically on the separation of the given energy level from the rest of the spectrum. Thus, the problem at hand is to run the Lanczos procedure within an invariant subspace. To be specific, let us consider the total spin S and its projection S_z.

In the case of S_z the solution is trivial, because S_z can be written as a sum of single-particle quantities (single-particle spin projections)

$$S_z = \sum_{i=1}^{N} s_z^{(i)}$$

and thus configurations contained in the basis can be chosen as the eigenstates of S_z. In this case, because of the conservation of S_z

$$\left[H, S_z\right] = 0$$

the Hamiltonian does not couple configurations of different S_z

$$\left\langle S_z \middle| H \middle| S_z' \right\rangle = \left\langle S_z \middle| H \middle| S_z \right\rangle \delta_{S_z S_z'}$$

This means that the Hamiltonian matrix is block-diagonal, and thus that each different block corresponding to a particular value of S_z can be diagonalized separately. Of course, this is easily done by generating the basis exclusively with configurations having the given S_z. Hence, resolution of the additional conserved quantum number S_z can be done already at the level of basis generation. It should be done whenever possible as (besides other benefits mentioned above) it obviously reduces dimension of the Hilbert space which is the key parameter determining speed and memory requirements of the algorithm.

The case of the total spin S cannot be solved in an analogous way, since the relevant operator S^2 is not a linear but quadratic expression in the single-particle operators

$$S^2 = \sum_{i=1}^{N} \left[\left(s_x^{(i)}\right)^2 + \left(s_y^{(i)}\right)^2 + \left(s_z^{(i)}\right)^2 \right] = \sum_{i=1}^{N} \left[s_+^{(i)} s_-^{(i)} + s_-^{(i)} s_+^{(i)} + s_z^{(i)} s_z^{(i)} \right]$$

and thus the (uncorrelated, by definition) single-particle configurations are not its eigenstates.

The first observation could be that if the initial Lanczos vector were chosen as an eigenvector of S, then (at least in the absence of numerical rounding errors) so would be each successive Lanczos vector – as a trivial consequence of the vanishing commutator

$$\left[H, S\right] = 0$$

and the whole Lanczos procedure would be automatically self-restricted to the invariant subspace of this initial vector.

How can an eigenstate with an arbitrary value of S be efficiently generated? In general, this is a problem of finding an eigenstate (of S^2) corresponding to a known eigenvalue (S(S+1)). In the case of total spin it can be further simplified by generating the basis with

$$S_z = S$$

and then looking for the solution to

$$S^- \psi = 0$$

which can be recast as looking for the (zero-mode) ground state of a Hermitian operator

$$S^+ S^-$$

(different from S^2 only by an S_z-dependent shift).

Formally, this is a two-particle operator which can be conveniently expressed in the second-quantization form as

$$S^+S^- = \sum_{ij} c^+_{i\uparrow} c^+_{j\downarrow} c_{i\downarrow} c_{j\uparrow}$$

and treated by an independent Lanczos procedure as if it were a kind of interaction (albeit evidently producing a much sparser matrix).

Once the initial Lanczos vector has been generated as an eigenvector of S, and the Lanczos algorithm is run with finite rounding errors, the same accumulation of errors as discussed earlier in the context of ghost states now emerges again. Namely, the successive Lanczos vectors are slowly losing their (initially full) projection onto the invariant subspace of S, so that after sufficiently many iterations the Lanczos procedure may return a ground state with a different S. However, this problem will only arise if the invariant subspace of S does not contain the actual ground (in which case the growing Lanczos basis will be automatically attracted to the true ground state, and thus leak out of the initial subspace corresponding to S). The accumulation of rounding errors is therefore harmless if the correct value of S for the true ground state can be anticipated, but when the goal is to find the true ground state anyway. When the sought energy spectrum at some value of S lies above a ground state at a different spin S_0, the accumulation of rounding errors must be remedied by, for example, a repeated projection onto the correct invariant S-subspace (at least after each number of iterations; in practice this is best done after each iteration). The projection onto the S-subspace is readily implemented as an independent Lanczos procedure for the operator

$$S^+S^-$$

started from the current Lanczos vector of the Lanczos procedure for H. In other words, the perpetual projection onto the S-subspace is achieved by nesting the Lanczos procedure for S^+S^- inside (after each iteration) the "main" Lanczos procedure for H. It is important to note that since a single Lanczos step for H can only cause a very small leakage of the next Lanczos basis vector outside of the S-subspace, such vector is guaranteed to be an almost exact ground state of S^+S^-. Hence, when fed into the internal S-Lanczos, it will usually converge quickly and produce a state that is virtually unchanged (only "straightened" in terms of S) so that the Lanczos connection within the main chain of vectors coupled by H remains essentially preserved (as is necessary for the main Lanczos procedure to converge to an eigenstate of H).

Let us finally remark that the small number of required iterations for the internal Lanczos loop is not always guaranteed by the near exactness of the initial vector for the internal procedure. Rather, it critically depends on the separation of the lowest (zero-mode) eigenstate of S^+S^- from the low-lying excitations. In the case of total N-electron spin, the spectrum contains the eigenvalues od the form S(S+1) which become progressively denser toward the lower end of the spectrum. The gap from the lowest state at

$$S(S+1) = 0$$

to the next state at

$$S(S+1) = 2$$

can become quite small compared to the width of the entire spectrum

$$S_{MAX}(S_{MAX}+1)$$

where

$$S_{MAX} = \frac{1}{2} N$$

when the system size N becomes large. To be precise

$$\frac{1 \cdot 2 - 0 \cdot 1}{S_{MAX}\left(S_{MAX}+1\right)} \propto \frac{1}{N^2}$$

which can cause somewhat slow convergence in large systems. This is even more problematic with the total orbital angular momentum which is conserved in systems with spherical symmetry. For example, in an atomic-like shell of angular momentum l the maximum total angular momentum of N fermions is

$$Nl - \frac{1}{2} N\left(N-1\right)$$

and the gap-to-width ratio becomes quickly even smaller than for the total spin. In those cases, running the nested Lanczos only after certain iterations of the main loop might be a better choice.

When more than one additional quantum number need be resolved (of course, assuming they all commute both with one another and with the Hamiltonian), this can be done efficiently by simply running the internal loops (for example, for S, L, …) one immediately after another. This is usually a more efficient option than seeking a zero mode of a linear combination

$$\alpha \cdot S^+ S^- + \beta \cdot L^+ L^-$$

with some optimized (positive) α and β. And definitely it is not advised to combine additional quantum numbers with the Hamiltonian in the form

$$H + \lambda \cdot S^+ S^-$$

in attempt to find its ground state for a very large λ (the convergence is then very poor due to the artificially widened spectrum – again, compared to the low-energy gap of H itself, which the Lanczos procedure is aimed at resolving).

8. Concrete example of configuration interaction method

A particular example of the application of the configuration interaction method comes from the field of fractional quantum Hall effect. The actual physical system of relevance is the two-dimensional electron gas in a quantizing (strong) perpendicular magnetic field B. 2D electron concentration ρ is such that the electrons fill a fraction ν of the lowest Landau level, whose macroscopic degeneracy g is proportional to the magnetic flux through the sample (of area A)

$$g = \frac{AB}{hc/e}$$

The single-particle cyclotron gap

$$\hbar\omega_c \propto B$$

exceeds the characteristic Coulomb energy

$$\frac{e^2}{\lambda} \propto \sqrt{B}$$

where λ is the magnetic length. In result, the electrons are effectively confined to an isolated, degenerate, partially filled (lowest) Landau level. After the constant total cyclotron energy is removed from the Hamiltonian by a trivial gauge transformation, all that is left is the electron-electron (Coulomb) interaction

$$H = \sum_{i,j} \frac{e^2}{r_{ij}}.$$

Perturbative treatment of the (sole) interaction term is not possible because of an obvious lack of a small expansion parameter, making this problem ideally suitable for an exact calculation. In order to employ the configuration interaction method, the actual extended system need be accurately modeled (represented) by a finite system of N electrons. To maintain the correct Landau level filling factor ν while using a finite electron number N, the area of the system A must also be restricted. This means applying some kind of boundary conditions. A particularly convenient geometry for the study of liquid many-electron phases (in contrast to the Wigner crystal or other broken-symmetry phases) was introduced by Haldane [26]. The so-called Haldane sphere has a Dirac magnetic monopole of strength

$$2Q\frac{hc}{e}$$

at its center, which produces an isotropic radial magnetic field, perpendicular to the surface and having a fixed magnitude at each point (on the surface). Although it has a finite area

$$4\pi R^2 = 4\pi Q\lambda^2$$

and thus also a finite Landau level degeneracy

$$g = 2Q + 1,$$

Haldane sphere is a highly symmetrical geometry, with the group of rotations replacing the group of translations of the original extended plane. Thus, the conservation laws which on a plane are a consequence of the invariance under the group of so-called magnetic translations translate into Haldane spherical geometry in form of the conservation of the total angular momentum L and its projection M. Indeed, there is a one-to-one correspondence between the conserved quantum numbers in both geometries, and the plane can be regarded as a limit of the spherical geometry (for an infinite radius when expressed in the units of magnetic length)

$$\frac{\lambda}{R} = \frac{1}{\sqrt{Q}} \to 0.$$

The first step in the configuration interaction is the solution of the single-particle problem [27]. Here, owing to the spherical symmetry, the single-particle states Y are obviously the eigenstates of the length *l* and projection *m* of the orbital angular momentum. They are called "monopole harmonics"

$$Y_{Q,l,m}\left(\vec{\Omega}\right)$$

and are generalizations of the spherical harmonics (the latter being the special case for Q=0).

The angular momenta allowed for an electron in the presence of a Dirac magnetic monopole of strength 2Q are

$$l_n = Q + n$$

with

$$n = 0, 1, \ldots$$

The energy only depends on angular momentum and reads

$$E_l = \frac{\hbar\omega_c}{2Q}\left[l(l+1) - Q^2\right].$$

The angular momentum shells represent degenerate Landau levels of the planar geometry, also separated by gaps which converge to the cyclotron energy in the limit of large 2Q (that is, of large R/λ), which is seen more clearly in the following (equivalent) expression

$$E_n = \hbar\omega_c\left[n + \frac{1}{2} + \frac{n(n+1)}{2Q}\right].$$

In particular, the lowest shell of n=0 represents the lowest Landau level.

Let us now turn to a particular correlated many-electron state called a Laughlin quantum liquid, which in an extended system occurs at the Landau level filling factor ν=1/3. An extended many-body wave-function on a plane can be uniquely mapped onto the Haldane sphere. In general, the relation between the filling factor ν, the Landau level degeneracy g, and the electron number N, which for an infinite system is simply

$$\nu = \frac{N}{g}$$

on a finite sphere involves an additional "shift" γ

$$2Q = \frac{N}{\nu} - \gamma$$

which depends on the particular form of the wave-function, but (for a given wave-function) not on the system size N. In particular, the Laughlin ν=1/3 liquid is represented by a series of finite non-degenerate (L=0) ground states on a sphere which occur at

$$2Q = 3(N-1).$$

So the problem to be addressed numericall is that of N electrons in a shell of angular momentum

$$l = \frac{3}{2}(N-1).$$

To be specific, let us choose N=6 and 2l=15. The single-particle states are labeled by the orbital angular momentum projection *m*. The allowed configurations are

$$\left|m_1, m_1, \ldots, m_6\right\rangle.$$

The two-body matrix elements

$$\langle m_1, m_2 | V | m_3, m_4 \rangle = \int d\vec{\Omega}_1 d\vec{\Omega}_2 \; Y^*_{Q,l,m_1}(\vec{\Omega}_1) Y^*_{Q,l,m_2}(\vec{\Omega}_2) \frac{e^2}{R|\vec{\Omega}_1 - \vec{\Omega}_2|} Y_{Q,l,m_3}(\vec{\Omega}_2) Y_{Q,l,m_4}(\vec{\Omega}_1)$$

can be derived analytically or computed [26]. They are related to the so-called Haldane pair pseudopotential V_L (pair energy as a function of relative pair angular momentum) through the Clebsch-Gordan coefficients

$$\langle m_1, m_2 | V | m_3, m_4 \rangle = \sum_L \langle m_1, m_2 | L \rangle V_L \langle L | m_3, m_4 \rangle .$$

The projection M of the total orbital angular momentum can be resolved at the stage of basis generation, while its length L need be resolved by nesting the Lanczos loops.

Let us now count the number of relevant six-electron configurations (dimension of the Hilbert space or of the Hamiltonian matrix to be diagonalized). For N distinguishable particles the number of states would be huge

$$D = g^N = 16^6 = 16\,777\,216$$

(recall that g=2Q+1). For N particles which are still distinguishable but have hard cores preventing any two of them to occupy the same spin-orbital (as required by Pauli exclusion principle), the dimension shrinks to

$$D = g(g-1)\ldots(g-N+1) = \frac{g!}{(g-N)!} = \frac{16!}{10!} = 5\,765\,760$$

For N indistinguishable particles with hard cores, such as fermions whose wave function must be symmetric under exchange of any pair of coordinates, the dimension of the Hilbert space shrinks dramatically to

$$D = \binom{g}{N} = \binom{16}{6} = 8\,008 .$$

Due to rotational symmetry of the Hamiltonian, the Wigner-Eckart theorem guarantees that the energy and all other scalar quantities (such as pair correlation functions, etc.) only depend on the length of the total angular momentum L, but not on its projection M. Therefore , the diagonalization need not be repeated for each M. However, the Hilbert space for a given M only contains the states with

$$L \geq M .$$

Hence, in order to have a guarantee that the Hilbert space contains the ground state whose angular momentum L in not known in advance, one has to consider M=0.

The M=0 subspace has the dimension of

$$D = 338 .$$

Hence, the exact computation of the entire spectrum for the (N,2Q)=(6,16) system requires D=338 Lanczos iterations for the Hamiltonian matrix of (an equal) dimension D=338. If the eigenvectors were computed along with the eigenvalues, the energy spectrum can then be

resolved into different angular momentum (L) channels by the calculation of the expectation values of

L^+L^-

in each eigenstate.

Alternatively, L can be resolved by nesting the Lanczos loops for H and for L^+L^-. The dimensions of the relevant Hilbert spaces would then be

$$D_M = 338, 332, 330, 319, 310, 293, \ldots \quad \text{for} \quad M = 0, 1, 2, 3, 4, 5, \ldots$$

and the numbers of allowed multiplets corresponding to different L, are

$$\Gamma_{L,M} = D_{M=L} - D_{M=L+1} = 6, 2, 11, 9, 17, \ldots \quad \text{for} \quad L = 0, 1, 2, 3, 4, \ldots$$

(In this example, Γ_{ML} would also be the required number of iterations, as these are too small numbers to reach convergence even for the extremal eigenstates sooner than completing the Lanczos procedure.)

Let us now look at the concrete numerical results.

The following Fig.1 displays the energy spectrum for the above example of N=6 electrons on a Haldane sphere, in the lowest angular momentum shell (that is, in the lowest Landau level) for the magnetic flux 2Q=15. The configuration basis corresponds to M=0, and has dimension of D=338. The Lanczos tridiagonal Hamiltonian matrix is diagonalized after each Lanczos iteration, and the resulting energies are shown with the blue dots. For this particular figure, the Lanczos procedure is carried out with full orthogonalization (each new Lanczos vector is explicitly orthogonalized to all preceding Lanczos vectors, not only to the last two).

Clearly, while the complete energy spectrum is obtained after performing the number of 338 iterations equal to the full dimension of the dimension of configuration basis D, the extremal eigenenergies converge much sooner. This is especially true of the ground state which (as it represents an incompressible Laughlin liquid) separated from the excitations by a fairly large gap. The states at slightly higher energies converge after a somewhat larger number of iterations, but also much sooner than the states from the middle of the band, which require completing the Lanczos procedure to the last D^{th} iteration to converge.

The closeup view of the first 30 iterations of the same computation is presented in the following Fig.2, which most clearly shows the contrast in the rate of convergence between the extremal and isolated eigenvalues (for example, the ground state with an energy gap) and the levels from inside the energy spectrum.

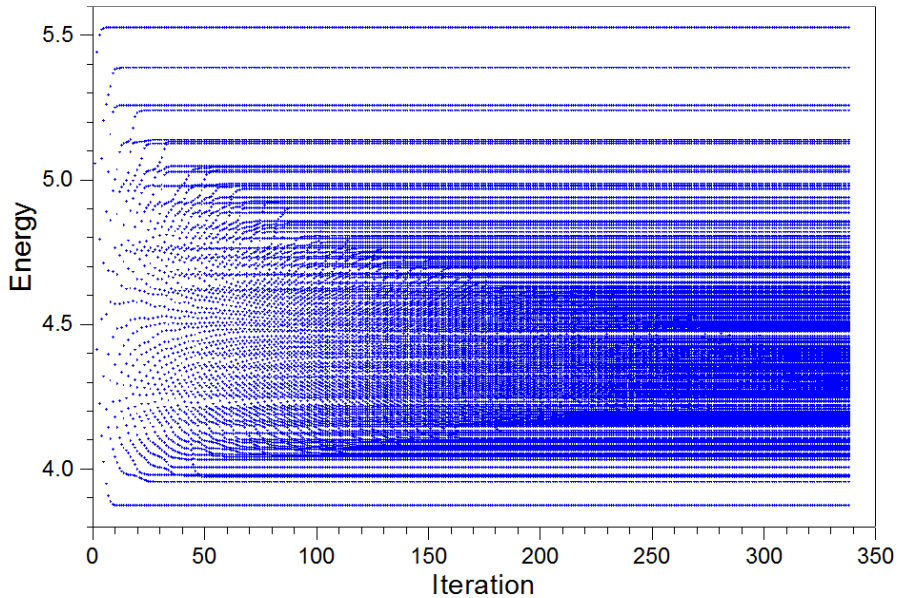


Fig.1 Full energy spectrum as a function of the number of Lanczos iterations, for the Lanczos algorithm carried out with full orthogonalization.

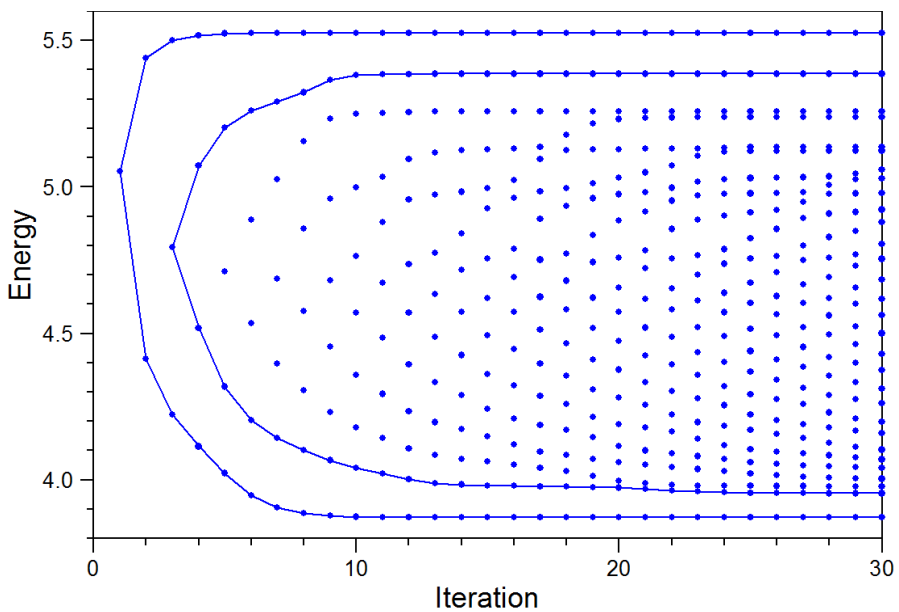


Fig.2 Close-up view of the low-iteration sector of Fig.1

The illustration of the emergence of "ghost states" is presented in the following Figs.3 and 4, which show an analogous evolution of the Lanczos energy spectrum as that of Figs.1 and 2, but this time for the Lanczos procedure carried out in a standard manner, that is with each new Lanczos vector explicitly orthogonalized to only two preceding ones. All other details are the same as in Fig.1 (in particular, the Hamiltonian matrix is identical).

At low energy (just above the ground state), the successive ghosts emerge from the above quasi-continuum and fall onto the ground state energy level at a fairly regular rate (around the iteration number 60, 105, 150, 190, 240, 280, and 320). As a result, the numerically computed ground state is 8-fold (nearly) degenerate after running D=338 iterations. On the other hand, the interior of the band has never reached convergence. And extracting reliable estimates of the excited energy levels (even of the few very lowest excitations) is problematic and requires a rather tedious analysis of the shown full evolution of the Lanczos spectrum. On the positive side, when the true degeneracy of the ground state is known in advance, the ground state energy (exclusively) converges just as quickly and reliably as in Fig.1, free of the ghost effect.

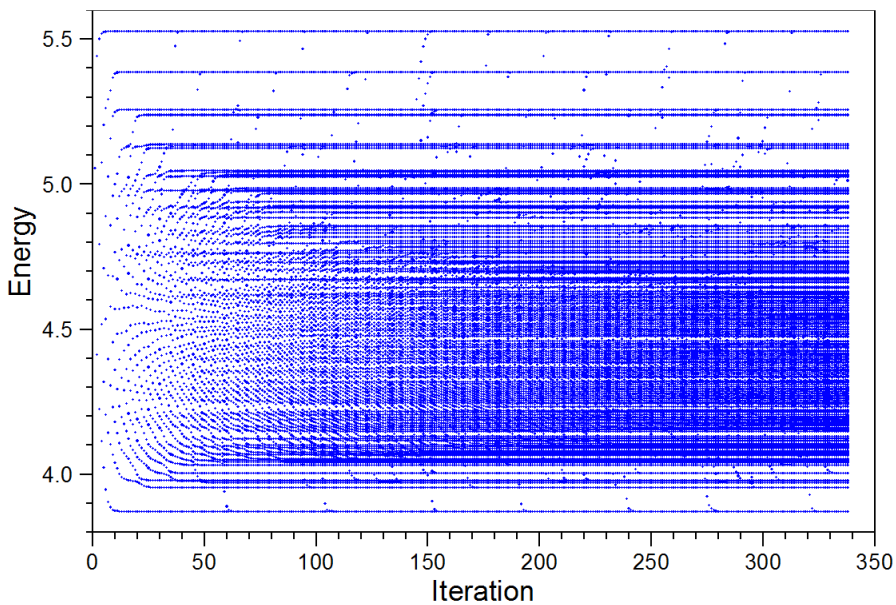


Fig.3 Full energy spectrum as a function of the number of Lanczos iterations, for the Lanczos algorithm carried out with orthogonalization performed to the previous two vectors.

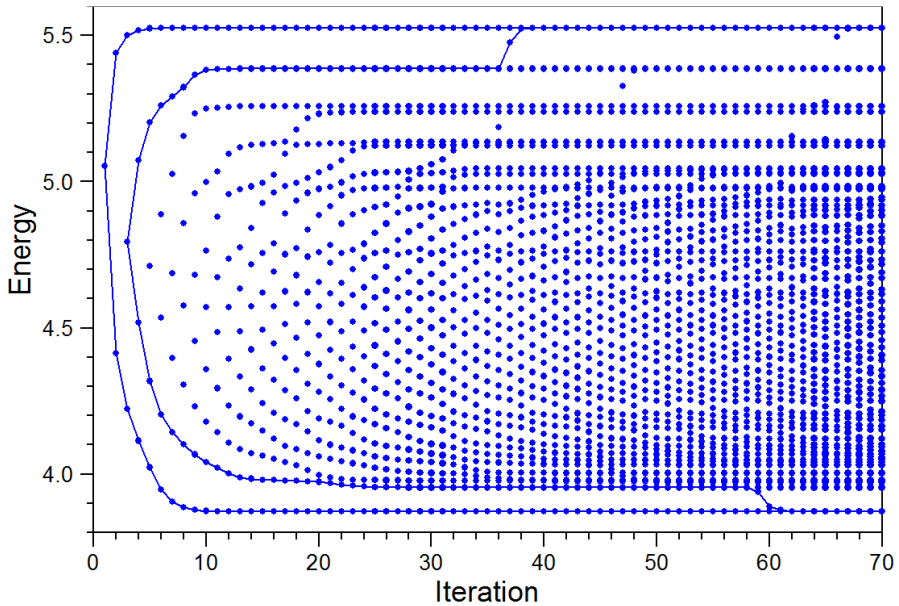


Fig.4 Zoom into the low-iteration sector of Fig.3

Finally, in Fig. 5 evolution of the Lanczos spectrum is shown for the calculation with the nested loops for the Hamiltonian and angular momentum, thus resolving the quantum number L in addition to M. This particular graph is for L=0. Thus, the Lanczos procedure is carried out in the same 338-dimensional configuration basis of M=0 as used for Figs. 1-4, but the complete set of Γ_0=6 eigenstates with L=0 is obtained after all Γ_0=6 iterations and, these six iterations being such a small number, all these iterations are indeed required to reach an accurate solution (even for the ground state).

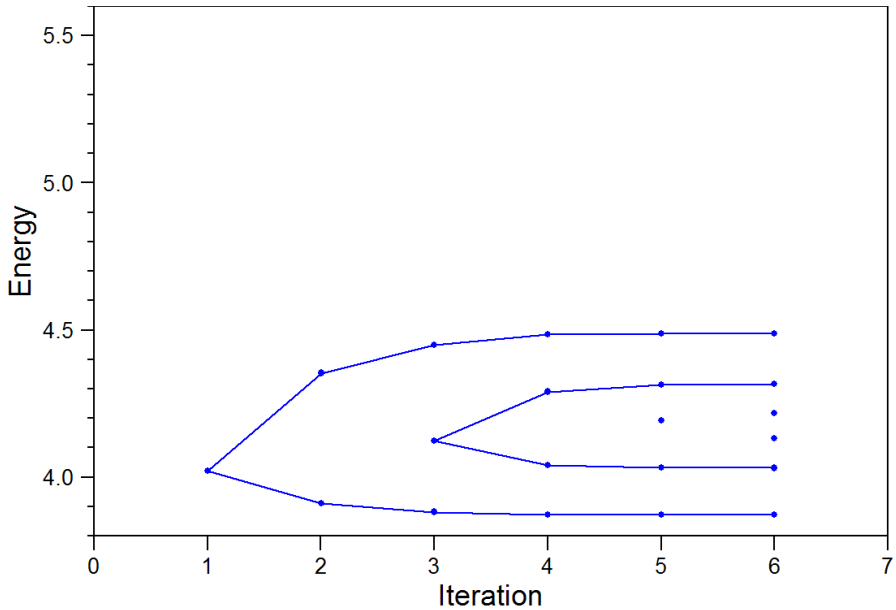


Fig.5 Energy spectrum in the zero total angular momentum (L=0) channel as a function of the number of Lanczos iterations, for the Lanczos algorithm for the squared angular momentum operator nested inside the Lanczos algorithm for the Hamiltonian.

Literature

[1] Reed M. A., Bate R. T., Bradshaw K., Duncan W. M., Frensley W. M., Lee J. W., Smith H. D., *Spatial quantization in GaAs-AlGaAs multiple quantum dots*, J. Vacuum Sci. Technol. B **4**, 358 (1986).

[2] Ashoori R. C., Störmer H. L., Weiner J. S., Pfeifer L. N., Pearton S. J., Baldwin K. W., West K., *Single-electron capacitance spectroscopy of discrete quantum levels*, Phys. Rev. Lett. **68**, 3088 (1992).

[3] Ashoori R. C., Störmer H. L., Weiner J. S., Pfeifer L. N., Baldwin K. W., West K., *N-electron ground state energies of a quantum dot in magnetic field*, Phys. Rev. Lett. **71**, 613 (1993).

[4] Hansen W., Smith T. P., Lee K. Y., Brum J. A., Knoedler C. M., Hong J. M., Kern D. P., *Zeeman bifurcation of quantum-dot spectra*, Phys. Rev. Lett. **62**, 2168 (1989).

[5] Lorke A., Kotthaus J. P., Ploog K., *Coupling of quantum dots on GaAs*, Phys. Rev. Lett. **64**, 2559 (1990).

[6] Meurer B., Heitmann D., Ploog K., *Single-electron charging of quantum-dot atoms*, Phys. Rev. Lett. **68**, 1371 (1992).

[7] Sikorski C., Merkt U., *Spectroscopy of electronic states in InSb quantum dots*, Phys. Rev. Lett. **62**, 2164 (1989).

[8] Alsmeier J., Batke E., Kotthaus J. P., *Voltage-tunable quantum dots on silicon*, Phys. Rev. B **41**, 1699 (1990).

[9] Hansen W., Smith T. P., Lee K. Y., Hong J. M., Knoedler C. M., *Fractional states in few-electron systems*, Appl. Phys. Lett. **56**, 168 (1990).

[10] Brunner K., Bockelmann U., Abstreiter G., Walther M., Böhm G., Tränkle G., Weimann G., *Photoluminescence from a~single GaAs/AlGaAs quantum dot*, Phys. Rev. Lett. **69**, 3216 (1992).

[11] Ekimov A. I., Efros A. L., Onushchenko A. A., *Quantum size effect in semiconductor microcrystals*, Solid State Commun. **56**, 921 (1985).

[12] Fukui T., Ando S., Tokura Y., *GaAs tetrahedral quantum dot structures fabricated using selective area metalorganic chemical vapor deposition*, Appl. Phys. Lett. **58**, 2018 (1991).

[13] Lebens J. A., Tsai C. S., Vahala K. J., *Application of selective epitaxy to fabrication of nanometer scale wire and dot structures*, Appl. Phys. Lett. **56**, 2642 (1990).

[14] Petroff P. M., Denbaars S. P., *MBE and MOCVD growth and properties of self-assembling quantum dot arrays in III-V semiconductor structures*, Superlattices and Microstructures **15**, 15 (1994).

[15] Raymond S., Fafard S., Poole P. J., Wójs A., Hawrylak P., Charbonneau S., Leonard D., Leon R., Petroff P. M., Merz J. L., *State filling and time-resolved photoluminescence of excited states in $In_{1-x}Ga_xAs/GaAs$ self-assembled quantum dots*, Phys. Rev. B **54**, 11548 (1996).

[16] Marzin J.-Y., Gerard J.-M., Izrael A., Barrier D., Bastard G., *Photoluminescence of single InAs quantum dots obtained by self-organized growth on GaAs*, Phys. Rev. Lett. **73**, 716 (1994).

[17] Leinaas J. M., Myrheim J., *On the theory of identical particles*, Il Nuovo Cimento B **37** 1 (1977).

[18] Wilczek F., *Quantum Mechanics of Fractional-Spin Particles*, Phys. Rev. Lett. **49**, 957 (1982)

[19] Tsui D. C., Stormer H. L., Gossard A. C., *Two-Dimensional Magnetotransport in the Extreme Quantum Limit*, Phys. Rev. Lett. **48**, 1559 (1982)

[20] Laughlin R. B., *Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations*, Phys. Rev. Lett. **50**, 1395 (1983).

[21] de-Picciotto R., Reznikov M., Heiblum M., Umansky V., Bunin G., Mahalu D., *Direct observation of a fractional charge*, Nature **389**, 162 (1997).

[22] Moore G., Read N., *Nonabelions in the fractional quantum hall effect*, Nucl. Phys. B **360**, 362 (1991).

[23] Hohenberg, P, Kohn W, *Inhomogeneous electron gas*, Phys. Rev. **136**, B864 (1964).

[24] Kohn W., Sham L. J., *Self-Consistent Equations Including Exchange and Correlation Effects*, Phys. Rev. **140**, A1133 (1965).

[25] Ceperley D. M., Alder B. J., *Ground State of the Electron Gas by a Stochastic Method*, Phys. Rev. Lett. **45**, 566 (1980).

[26] Haldane F. D. M., *Fractional Quantization of the Hall Effect: A Hierarchy of Incompressible Quantum Fluid States*, Phys. Rev. Lett. **51**, 605 (1983).

[27] Wu, T. T., Yang, C. N., *Dirac monopole without strings: monopole harmonics*, Nucl. Phys. B **107**, 365 (1976).